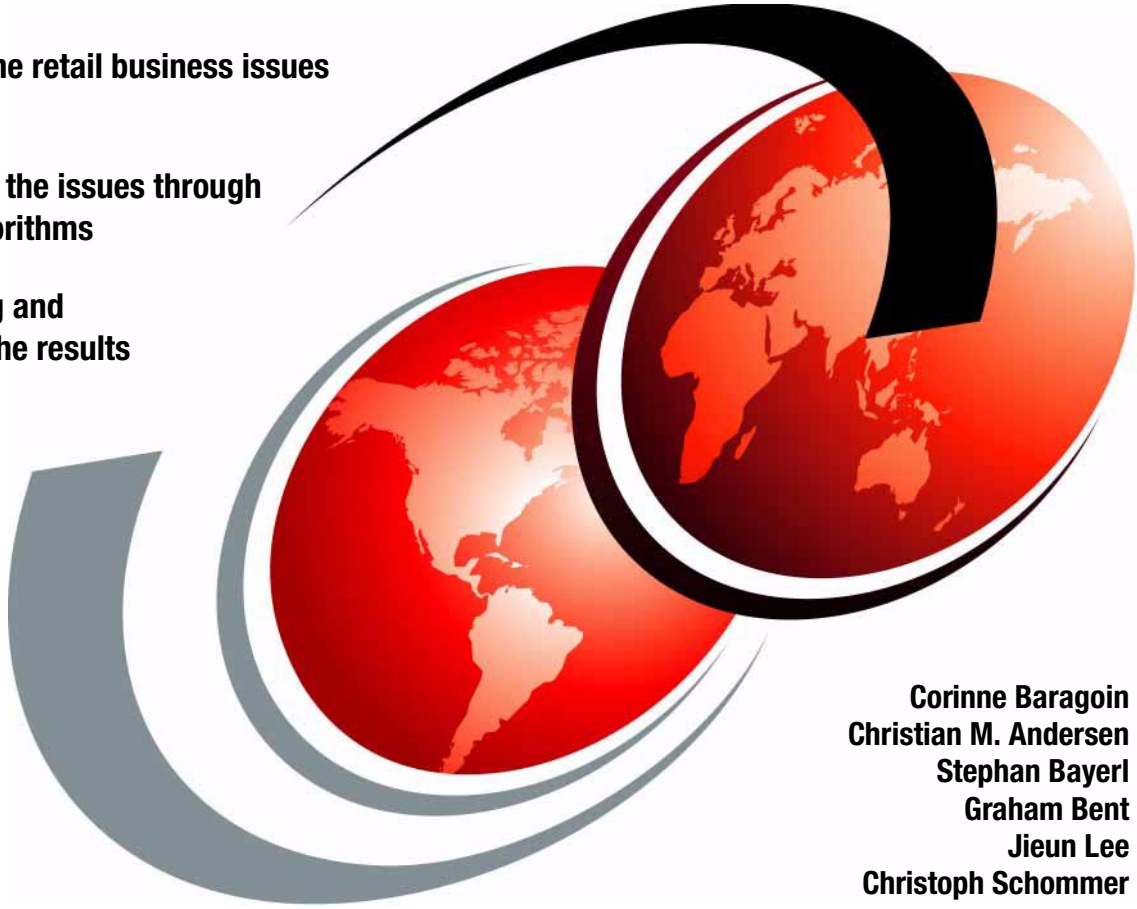
IBM

Mining Your Own Business in Retail

Using DB2 Intelligent Miner for Data

- Exploring the retail business issues
- Addressing the issues through mining algorithms
- Interpreting and deploying the results



Corinne Baragoin
Christian M. Andersen
Stephan Bayerl
Graham Bent
Jieun Lee
Christoph Schommer

Redbooks

ibm.com/redbooks

International Technical Support Organization

Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data

August 2001

SG24-6271-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 187.

First Edition (August 2001)

This edition applies to IBM Intelligent Miner For Data V6.1.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Preface

The data you collect about your customers, is one of the greatest assets that any business has available. Buried within the data is all sorts of valuable information that could make a significant difference to the way you run your business and interact with your customers. But how can you discover it?

This IBM Redbook focuses on a specific industry sector, the retail sector, and explains how IBM DB2 Intelligent Miner For Data (IM for Data) is the solution that will allow you to mine your own business.

This redbook is one of a family of redbooks that has been designed to address the types of business issues that can be solved by data mining in different industry sectors. The other redbooks address the banking, telecoms and health care sectors.

Using specific examples for retail, this book will help business decision makers to understand the sorts of business issues that data mining can address, how to interpret the mining results, and how to deploy them in the business. Business decision makers will want to skip certain sections of the book, such as “The data to be used”, “Sourcing and preprocessing the data”, and “Evaluating the data”.

This book will also help implementers to understand how a generic mining method can be applied. This generic method describes how to translate the business issues into a data mining problem and some common data models that you can use. It explains how to choose the appropriate data mining technique and then how to interpret and deploy the results in the enterprise.

Although no in-depth knowledge of Intelligent Miner For Data is required, a basic understanding of data mining technology is assumed.

The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Corinne Baragoin is a Business Intelligence Project Leader at the International Technical Support Organization, San Jose Center. Before joining the ITSO, she had been working as an IT Specialist for IBM France, assisting customers on DB2 and data warehouse environments.

Christian M. Andersen is a Business Intelligence/CRM Consultant for IBM Nordics. He holds a degree in Economics from the University of Copenhagen. He has many years of experience in the data mining and business intelligence field. His areas of expertise include business intelligence and CRM architecture and design, spanning the entire IBM product and solution portfolio.

Stephan Bayerl is a Senior Consultant at the IBM Boeblingen Development Laboratory in Germany. He has over four years of experience in the development of data mining and more than three years in applying data mining to business intelligence applications. He holds a doctorate in Philosophy from Munich University. His other areas of expertise are in artificial intelligence, logic, and linguistics. He is a member of Munich University, where he gives lectures in analytical philosophy.

Graham Bent is a Senior Technology Leader at the IBM Hursley Development Laboratory in the United Kingdom. He has over 10 years of experience in applying data mining to military and civilian business intelligence applications. He holds an master's degree in Physics from Imperial College (London) and a doctorate from Cranfield University. His other areas of expertise are in data fusion and artificial intelligence.

Jieun Lee is an IT Specialist for IBM Korea. She has five years of experience in the business intelligence field. She holds a master's degree in Computer Science from George Washington University. Her areas of expertise include data mining and data management in business intelligence and CRM solutions.

Christoph Schommer is a Business Intelligence Consultant for IBM Germany. He has five years of experience in the data mining field. He holds a master's degree in Computer Science from the University of Saarbruecken and a doctorate of Health care from the Johann Wolfgang Goethe-University Frankfurt in Main, Germany. His areas of expertise include the application of data mining in different industrial areas. He has written extensively on the application of data mining in practice.

Thanks to the following people for their contributions to this project:

- By providing their technical input and valuable information to be incorporated within these pages:

  Frank Theisen
  Gregor Meyer
  Mahendran Maliapen
  Martin Brown
  IBM

- By answering technical questions and reviewing this redbook:

  Andreas Arning
  Christoph Lingenfelder
  Reinhold Keuler
  Ute Baumbach
  Intelligent Miner Development Team at the IBM Development Lab in Boeblingen

- By reviewing this redbook:

  Gerd Piel
  Jim Lyon
  Richard Hale
  Steve Addison
  Tom Bradshaw
  IBM

Special notice

This publication is intended to help business decision makers to understand the sorts of business issues that data mining can address and to help implementers, starting with data mining, to understand how a generic mining method can be applied. The information in this publication is not intended as the specification of any programming interfaces that are provided by IBM DB2 Intelligent Miner For Data. See the PUBLICATIONS section of the IBM Programming Announcement for IBM DB2 Intelligent Miner For Data for more information about what publications are considered to be product documentation.

IBM trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

e (logo)® 
IBM ®
AIX
AT
CT
Current
DataJoiner
Redbooks
Redbooks Logo 
DB2
DB2 Universal Database
Information Warehouse
Intelligent Miner
SP

Comments welcome

Your comments are important to us!

We want our IBM Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to the address on page ii.

1

Introduction

In today's dynamic business environment, successful organizations must be able to react rapidly to the changing market demands.

To do this requires an understanding of all of the factors that have an influence on your business, and this in turn requires an ability to monitor these factors and provide the relevant and timely information to the appropriate decision makers.

Creating a picture of what is happening relies on the collection, storage, processing and continuous analysis of large amounts of data to provide the information that you need. This whole process is what we call Business Intelligence (BI). BI is about making well-informed decisions, using information that is based on data. Data in itself provides no judgement or interpretation and therefore provides no basis for action. Putting data into context is what turns it into information. Connecting pieces of available information leads to the knowledge that can be used to support decisions. Where the context is well understood, BI enables the transformation from data to decision to become a routine process within your business. One of the main challenges is that increasing competitive pressures requires new and innovative ways to satisfy increasing customer demands. In these cases the context is not well understood.

Data mining provides the tools and techniques to help you *discover* new contexts and hence new things about your customers. Mining your own business will enable you to make decisions based upon real knowledge instead of just a gut feeling.

1.1 Why you should mine your own business

Increasing competitive pressures require you to develop new and innovative ways to satisfy the increasing demands your customers make. To develop these new ideas requires information about your customers and this information in turn must be derived from the data you collect about your customers. This information is not only invaluable from the perspective of your own business but is also of interest to the suppliers who manage the brands that you sell. Your data should be seen as one of the greatest assets your business owns.

The challenge that faces most retail organizations is that the volumes of data that can potentially be collected are so huge and the range of customer behavior is so diverse that it seems impossible to rationalize what is happening. If you are reading this book and you don't mind being told to "mine your own business" then you are probably already in this position. The question we want to address is, how can data mining help you discover new things about your customers and how can you use this information to drive your business forward?

The road from data to information, and finally to the decision making process itself, is not an easy one. In this book our objective is to show, through some example cases, what role data mining has to play in this process, what sorts of retail business problems you can address and what you need to do to mine your own business and to improve your Customer Relationship Management (CRM).

1.2 What are the retail business issues to address?

There are a large number of retail business questions to which data mining can provide answers, for example:

- What are the characteristics of my customers?
- What mix of products should I have?
- To which customers should I target specific products or offers?
- How should the products be placed within my retail outlet?
- Where should I place new stores?

The data mining techniques that we can use to obtain these answers are the subject of this book. It would take much larger book than this one to address all of the questions that data mining can answer, and therefore we have chosen to restrict ourselves to just three specific business issues. Our choice of issues have been selected primarily to illustrate the range of the data mining techniques that we have available to us.

So in this book we look at answering three specific questions to illustrate how these techniques can be applied:

- How can I characterize my customer from the mix of products that they purchase?
- How can I categorize my customers and identify new potential customers?
- How can I decide which products to recommend to my customers?

In our first example we consider the question of *how to characterize your customers* using data that you routinely collect about them. This example illustrates how you can discover customer segments from your data rather than having to use subjective business rules to identify the different types of customers you have. To generate the segments we use the data mining technique of c*lustering* and we will show you how you can use this technique to map out where your different customers are in relation to each other. The results that we produce can be used, for example, to identify the different types of customer that you have, potential niche market segments and where opportunities exist to cross sell. We will also explain how you can operationally deploy the results into your business.

Where our first example looks at how to discover new segments or categories of customers, in the second example we consider the question of *how to classify your customers* into categories that you have previously defined. The data mining technique that we use to do this is called c*lassification*. As we will show, these techniques have many potential applications in the retail industry, for example, identifying potential new profitable customers. In our specific example we will show you how to develop classification models that can be deployed around your business to classify customers at the point of sale, or in kiosks or other customer touch points. We will also explain how classification can be used to provide the vital information required for targeted marketing campaigns.

In our final example we consider the question of *how to cross-sell and up-sell products to your customers*. We will show you how you can use data mining to identify these opportunities using a combination of the clustering data mining techniques and a technique known as a*ssociations mining or market basket analysis*. The results of this type of data mining can be used, for example, to determine which types of products should be placed together in your retail outlet or, as we will show, to produce automatic product recommendation systems.

By concentrating on these three questions we hope that you will be able to appreciate why you should mine your own business with the ultimate objective of deploying the results of the data mining into your business process.

1.3 How this book is structured

The main objective of this book is to address the above retail business issues using data mining techniques.

However, to put this into context, it is first necessary to understand the context of data mining in an overall BI architecture. The road from data to decisions is not an easy one, and if you are going to mine your own business you will need some guidance.

To help in both of these areas:

- Chapter 2, “Business Intelligence architecture overview” on page 7 provides a BI architecture overview.
- Chapter 3, “A generic data mining method” on page 23 presents a detailed overview of what data mining describes as a generic method that can be followed.
- For the examples in the following chapters, use these methods and apply them to the business questions:
  - Chapter 4, “How can I characterize my customers from the mix of products that they purchase?” on page 45.
  - Chapter 5, “How can I categorize my customers and identify new potential customers?” on page 97.
  - Chapter 6, “How can I decide which products to recommend to my customers?” on page 137.
- Finally, in Chapter 7, “The value of DB2 Intelligent Miner For Data” on page 171, we describe the benefits of IM for Data, the data mining tool that we use in these examples.

We have provided sufficient information for you to understand how you are able to mine your own retail business without going into too many technical details about the data mining algorithms themselves. There is a difficult balance to strike here, and therefore for you to decide which sections you should read, we want to make the following comments:

We do **not:**

- Provide a user’s guide of any mining function
- Explicate any mining function in a mathematical complete way
- Deliver the basic background knowledge of a statistical introductory book
- Stress a particular data mining toolkit
- Provide a comparison of competitive mining products

Rather, we stress an operational approach to data mining by explaining the:

- Mechanics of operating a data mining toolkit
- Generic method as a guideline for the newcomer and the expert
- Technical aspects of the mining algorithms
- Necessary data preparation steps in a detailed manner
- Proven mining applications in the field
- Further steps for improvement of the mining results

It is assumed that a task like ours must remain incomplete in the sense that all examples demonstrated in this book could be copied and exploited in a short time, from several days to some weeks, while serious mining projects run from several weeks to months and longer. Therefore, the book lacks the description of the necessary bothersome and tedious mining cycles and does not offer a list of helpful tricks to simplify or overcome them totally. And of course, the approaches presented here by no means embrace all types of business issues.

1.4 Who should read this book?

This book is intended:

- To help business users figure out how data mining can address and solve specific business issues by reading the following sections in the different chapters:
  - The business issue
  - Interpreting the results
  - Deploying the mining results
- To be a guide for implementers on how to use data mining to solve business issues by explaining and detailing the generic method in each business question chapter, by providing data models to use and by including some experience-based hints and tips. It is worthwhile for implementers to progress sequentially through each business question chapter.
- To provide additional information:
  - To position data mining in the business intelligence architecture by reading Chapter 2, “Business Intelligence architecture overview” on page 7
  - To evaluate the data mining product by reading Chapter 7, “The value of DB2 Intelligent Miner For Data” on page 171

To benefit from this book, the reader should have, at least, a basic understanding of data mining.

2

Business Intelligence architecture overview

Business Intelligence (BI) covers the process of transforming data from your various data sources into meaningful information that can provide you and your company with insights into where your business has been, is today, and is likely to be tomorrow.

BI allows you to improve your decision-making at all levels by giving you a consistent, valid, and in-depth view of your business by consolidating data from different systems into a single accessible source of information — a data warehouse.

Depending on the users' needs there are different types of tools to be used to analyze and visualize the data from the data warehouse. These tools range from query and reporting to advanced analysis by data mining.

In this chapter we will describe the different components in a BI architecture. This will lead you to an overview of the architecture on which your data mining environment will be founded.

2.1 Business Intelligence

Traditionally, information systems have been designed to process discrete transactions in order to automate tasks, such as order entry, or account transactions. These systems however are not designed to support users who wish to extract data at different aggregation levels and utilize advanced methods for data analysis. Apart from these, systems tend to be isolated to support a single business system. This results in a great challenge when requiring a consolidated view of the state of your business.

This is where data warehouse and analytical tools come to your aid.

2.2 Data warehouse

Figure 2-1 shows the entire data warehouse architecture in a single view. The following sections will concentrate on single parts of this architecture and explain them in detail.

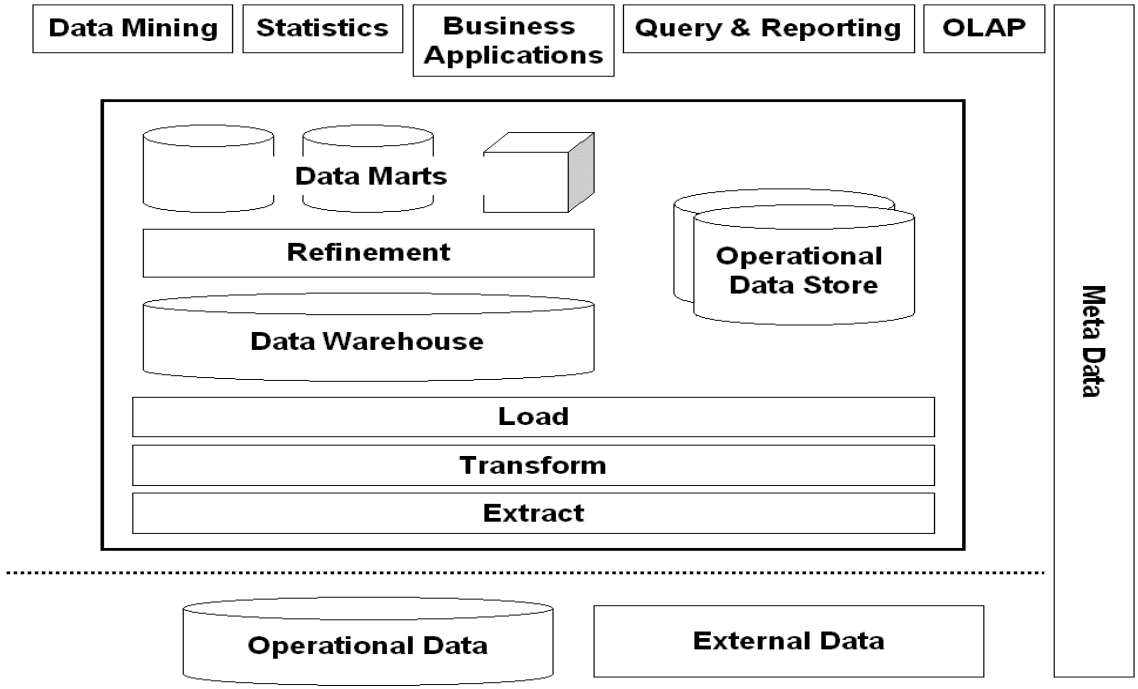


Figure 2-1 Data warehouse components

The processes required to keep the data warehouse up to date as marked are:

- Extraction/propagation
- Transformation/cleansing
- Data refining
- Presentation
- Analysis tools

The different stages of aggregation in the data are:

- On Line Transaction Programs (OLTP) data
- Operational Data Store (ODS)
- Datamarts

Metadata and how it is involved in each process is shown with solid connectors.

The tasks to be performed on the dedicated OLTP system are optimized for interactive performance and to handle the transaction oriented tasks in the day-to-day-business.

The tasks to be performed on the dedicated data warehouse machine require high batch performance to handle the numerous aggregation, pre calculation, and query tasks.

2.2.1 Data sources

Data sources can be operational databases, historical data (usually archived on tapes), external data (for example, from market research companies or from the Internet), or information from the already existing data warehouse environment. The data sources can be relational databases from the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plain text files or pictures and other multimedia information.

2.2.2 Extraction/propagation

Data extraction / data propagation is the process of collecting data from various sources and different platforms to move it into the data warehouse. Data extraction in a data warehouse environment is a selective process to import decision-relevant information into the data warehouse.

Data extraction / data propagation is much more than mirroring or copying data from one database system to another. Depending on the technique, this process is either referred as:

- **Pulling** (Extraction of data)

  Or

- **Pushing** (Propagation of data)

2.2.3 Transformation/cleansing

Transformation of data usually involves code resolution with mapping tables, for example, changing the variable *gender* to:

- *0* if the value is *female*
- *1* if the value is *male*

It involves changing the resolution of hidden business rules in data fields, such as account numbers. Also the structure and the relationships of the data are adjusted to the analysis domain. Transformations occur throughout the population process, usually in more than one step. In the early stages of the process, the transformations are used more to consolidate the data from different sources; whereas, in the later stages, data is transformed to satisfy a specific analysis problem and/or a tool requirement.

Data warehousing turns data into information; on the other hand, **data cleansing** ensures that the data warehouse will have valid, useful, and meaningful information. Data cleansing can also be described as standardization of data. Through careful review of the data contents, the following criteria are matched:

- Replace missing values
- Normalize value ranges and units (for example, sales in the euro or dollar)
- Use valid data codes and abbreviations
- Use consistent and standard representation of the data
- Use domestic and international addresses
- Consolidate data (one view), such as house holding

2.2.4 Data refining

The atomic level of information from the star schema needs to be aggregated, summarized, and modified for specific requirements. This data refining process generates datamarts that:

- Create a subset of the data in the star schema
- Create calculated or virtual fields
- Summarize the information
- Aggregate the information

The layer in the data warehouse architecture is needed to increase the query performance and minimize the amount of data that is transmitted over the network to the end user query or analysis tool.

When talking about data transformation/cleansing, there are basically two different ways where the result is achieved. In detail, these are:

- **Data aggregation**: Changes the level of granularity in the information.

  Example: The original data is stored on a daily basis — the data mart contains only weekly values. Therefore, data aggregation results in less records.

- **Data summarization**: Adds up values in a certain group of information.

  Example: The data refining process generates records that contain the revenue of a specific product group, resulting in more records.

Data preparation for mining is usually a very time consuming task, often the mining itself requires less effort. The optimal way to do data preprocessing for data mining is typically very dependent on the technology used and the current skills, the volume of data to be processed and the frequency of updates.

2.2.5 Datamarts

Figure 2-1 shows where datamarts are located logically within the BI architecture.

A datamart contains data from the data warehouse tailored to support the specific requirements of a given business unit, business function or application.

The main purpose of a data mart can be defined as follows:

- To store pre-aggregated information
- To control end user access to the information
- To provide fast access to information for specific analytical needs or user group
- To represent the end users view and data interface of the data warehouse
- To create the multidimensional/relational view of the data

The database format can either be multidimensional or relational.

When building data marts, it is important to keep the following in mind:

- Data marts should always be implemented as an extension of the data warehouse, not as an alternative. All data residing in the data mart should

therefore also reside in the data warehouse. In this way the consistency and reuse of data is optimized.

- Data marts are typically constructed to fit one requirement, ideally. However, you should be aware of the trade-off between the simplicity of design (and performance benefits) compared to the cost of administrating and maintaining a large number of data marts.

2.2.6 Metadata

The metadata structures the information in the data warehouse in categories, topics, groups, hierarchies and so on. They are used to provide information about the data within a data warehouse, as given in the following list (also see Figure 2-2):

- Metadata are “subject oriented” and are based on abstractions of real-world entities, for example, “project”, “customer”, or “organization”.
- Metadata define the way in which the transformed data is to be interpreted, for example, “5/9/99” = 5th September 1999 or 9th May 1999 — British or US?
- Metadata give information about related data in the data warehouse.
- Metadata estimate response time by showing the number of records to be processed in a query.
- Metadata hold calculated fields and pre-calculated formulas to avoid misinterpretation, and contain historical changes of a view.

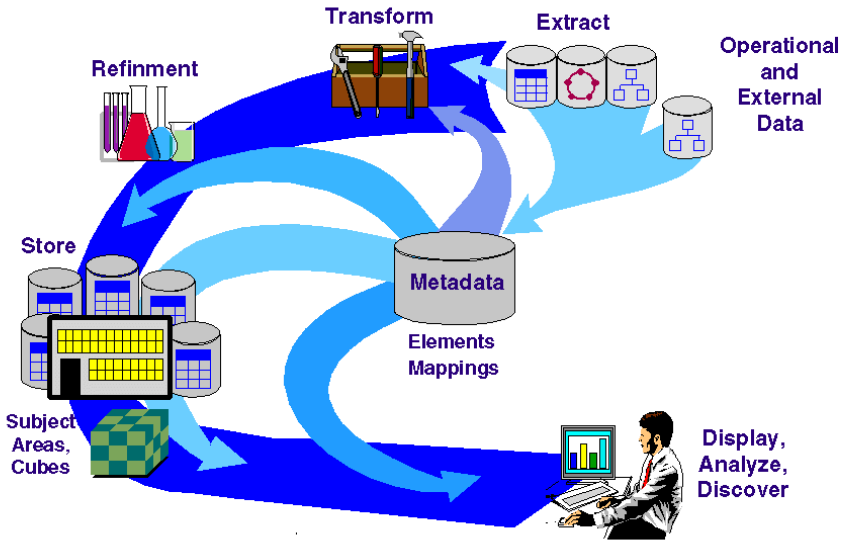


Figure 2-2 Metadata with a central role in BI

The data warehouse administrator's perspective of metadata is a full **repository** and documentation of all contents and processes within the data warehouse; from an end user perspective, metadata is the **roadmap** through the information in the data warehouse.

Technical versus business metadata

Metadata users can be broadly placed into the categories of business users and technical users. Both of these groups contain a wide variety of users of the data warehouse metadata. They all need metadata to identify and effectively use the information in the data warehouse.

Therefore, we can distinguish between two types of metadata that the repository will contain technical and business metadata:

- Technical metadata
- Business metadata

Technical metadata provides accurate data in the data warehouse. In addition, technical metadata is absolutely critical for the ongoing maintenance and growth of the data warehouse. Without technical metadata, the task of analyzing and implementing changes to a decision support system is significantly more difficult and time consuming.

The business metadata is the link between the data warehouse and the business users. Business metadata provides these users with a road map for access to the data in the data warehouse and its datamarts. The business users are primarily executives or business analysts and tend to be less technical; therefore, they need to have the DSS system defined for them in business terms. The business metadata presents, in business terms, what reports, queries and data are in the data warehouse; location of the data; reliability of the data; context of the data, what transformation rules were applied; and from which legacy systems the data was sourced.

Types of metadata sources

There are two broad types of metadata sources — formal and informal metadata. These sources comprise the business and technical metadata for an organization.

- Formal metadata sources are those sources of metadata that have been discussed, documented and agreed upon by the decision-makers of the enterprise. Formal metadata is commonly stored in tools or documents that are maintained, distributed and recognized throughout the organization. These formal metadata sources populate both technical and business metadata.
- Informal metadata consist of corporate knowledge, policies and guidelines that are not in a standard form. This is the information that people already know. This type of information is located in the “company consciousness” or it could be on a note on a key employee's desk. It is not formally documented or agreed upon; however, this knowledge is every bit as valuable as that in the formal metadata sources. Often, informal metadata provides some of the most valuable information, because it tends to be business related. It is important to note that in many cases much of the business metadata is really informal. As a result, it is critical that this metadata is captured, documented, formalized and reflected in the data warehouse. By doing this you are taking an informal source of metadata and transforming it into a formal source. Since every organization differs, it is difficult to say where your informal sources of metadata are; however, the following is a list of the most common types of informal metadata:
  - Data stewardship
  - Business rules
  - Business definitions
  - Competitor product lists

2.2.7 Operational Data Store (ODS)

The operational data source can be defined as an updateable set of integrated data used for enterprise-wide tactical decision making. It contains live data, not snapshots, and has minimal history that is retained. Below are some features of an Operational Data Store (ODS):

- An ODS is **subject oriented:** It is designed and organized around the major data subjects of a corporation, such as “customer” or “product”. They are not organized around specific applications or functions, such as “order entry” or “accounts receivable”.
- An ODS is **integrated:** It represents a collectively integrated image of subject-oriented data which is pulled in from potentially any operational system. If the “customer” subject is included, then all of the “customer” information in the enterprise is considered as part of the ODS.
- An ODS is **current valued:** It reflects the “current” content of its legacy source systems. “Current” may be defined in various ways for different ODSs depending on the requirements of the implementation. An ODS should not contain multiple snapshots of whatever “current” is defined to be. That is, if “current” means one accounting period, then the ODS does not include more than one accounting period’s data. The history is either archived or brought into the data warehouse for analysis.
- An ODS is **volatile:** Because an ODS is current valued, it is subject to change on a frequency that supports the definition of “current.” That is, it is updated to reflect the systems that feed it in the true OLTP sense. Therefore, identical queries made at different times will likely yield different results, because the data has changed.
- An ODS is **detailed:** The definition of “detailed” also depends on the business problem that is being solved by the ODS. The granularity of data in the ODS may or may not be the same as that of its source operational systems.

The features of an ODS such as subject oriented, integrated and detailed could make it very suitable to mining. These features alone do not make an ODS a good source for mining/training, because there is not enough history information.

2.3 Analytical users requirements

From the end user’s perspective, the presentation and analysis layer is the most important component in the BI architecture.

Depending on the user’s role in the business, their requirements for information and analysis capabilities will differ. Typically, the following user types are present in a business:

- The “non-frequent user”
  - This user group consists of people who are not interested in data warehouse details but have a requirement to get access to the information from time to time. These users are usually involved in the day-to-day business and do not have time or any requirements to work extensively with the information in the data warehouse. Their virtuosity in handling reporting and analysis tools is limited.
- Users requiring up-to-date information in predefined reports
  - This user group has a specific interest in retrieving precisely defined numbers in a given time interval, such as:

    “I have to get this quality-summary report every Friday at 10:00 AM as preparation to our weekly meeting and for documentation purposes.”
- Users requiring dynamic or ad hoc query and analysis capabilities
  - Typically, this is the business analyst. All the information in the data warehouse might be of importance to these users, at some point in time. Their focus is related to availability, performance, and drill-down capabilities to “slice and dice” through the data from different perspectives at any time.
- The advanced business analyst — the “power user”
  - This is a professional business analyst. All the data from the data warehouse is potentially important to these users. They typically require separate specialized datamarts for doing specialized analysis on preprocessed data. Examples of these are data mining analysts and advanced OLAP users.

Different user-types need different front-end tools, but all can access the same data warehouse architecture. Also, the different skill levels require a different visualization of the result, such as graphics for a high-level presentation or tables for further analysis.

In the remainder of this chapter we introduce the different types of tools that are typically used to leverage the information in a data warehouse.

2.3.1 Reporting and query

Creating reports is a traditional way of distributing information in an organization. Reporting is typically static figures and tables that are produced and distributed with regular time intervals or for a specific request. Using an automatic reporting tool is an efficient way of distributing the information in your data warehouse through the Web or e-mails to the large number of users, internal or external to your company, that will benefit from information.

Users that require the ability to create their own reports on the fly or wish to elaborate on the data in existing reports will use a combined querying and reporting tool. By allowing business users to design their own reports and queries, a big workload from an analysis department can be removed and valuable information can become accessible to a large number of (non-technical) employees and customers resulting in business benefit for your company. In contrast to traditional reporting this also allows your business users to always have access to up-to-date information about your business. This thereby also enables them to provide quick answers to customer questions.

As the reports are based on the data in your data warehouse they supply a 360 degree view of your company's interaction with its customers by combining data from multiple data sources. An example of this is the review of a client's history by combining data from: ordering, shipping, invoicing, payment, and support history.

Query and reporting tools are typically based on data in relational databases and are not optimized to deliver the "speed of thought" answers to complex queries on large amounts of data that is required by advanced analysts. An OLAP tool will allow this functionality at the cost of increased load time and management effort.

2.3.2 On-Line Analytical Processing (OLAP)

During the last ten years, a significant percentage of corporate data has migrated to relational databases. Relational databases have been used heavily in the areas of operations and control, with a particular emphasis on transaction processing (for example, manufacturing process control, brokerage trading). To be successful in this arena, relational database vendors place a premium on the highly efficient execution of a large number of small transactions and near fault tolerant availability of data.

More recently, relational database vendors have also sold their databases as tools for building data warehouses. A data warehouse stores tactical information that answers "who?" and "what?" questions about past events. A typical query submitted to a data warehouse is: "What was the total revenue for the eastern region in the third quarter?"

It is important to distinguish between the capabilities of a data warehouse from those of an On-Line Analytical Processing (OLAP) system. In contrast to a data warehouse — that is usually based on relational technology — OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis.

OLAP enables analysts, managers, and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information. OLAP transforms raw data so that it reflects the real dimensionality of the enterprise as understood by the user.

While OLAP systems have the ability to answer “who?” and “what?” questions, it is their ability to answer “what if?” and “why?” that sets them apart from data warehouses. OLAP enables decision making about future actions.

A typical OLAP calculation is more complex than simply summing data, for example: “What would be the effect on soft drink costs to distributors if syrup prices went up by $.10/gallon and transportation costs went down by $.05/mile?”

OLAP and data warehouses are complementary. A data warehouse stores and manages data. OLAP transforms data warehouse data into strategic information.

OLAP ranges from basic navigation and browsing (often known as “slice” and “dice”) to calculations, to more serious analyses, such as time series and complex modeling. As decision makers exercise more advanced OLAP capabilities, they move from data access to information to knowledge.

Who uses OLAP and why?

OLAP applications span a variety of organizational functions. Finance departments use OLAP for applications, such as budgeting, activity-based costing (allocations), financial performance analysis, and financial modeling. Sales analysis and forecasting are two of the OLAP applications found in sales departments. Among other applications, marketing departments use OLAP for market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation. Typical manufacturing OLAP applications include production planning and defect analysis.

Important to all of the above applications is the ability to provide managers with the information they need to make effective decisions about an organization's strategic directions. The key indicator of a successful OLAP application is its ability to provide information as needed, that is, its ability to provide “just-in-time” information for effective decision-making. This requires more than a base level of detailed data.

Just-in-time information is computed data that usually reflects complex relationships and is often calculated on the fly. Analyzing and modeling complex relationships are practical only if response times are consistently short. In addition, because the nature of data relationships may not be known in advance, the data model must be flexible. A truly flexible data model ensures that OLAP systems can respond to changing business requirements as needed for effective decision making.

Although OLAP applications are found in widely divergent functional areas, they all require the following key features:

- Multidimensional views of data
- Calculation-intensive capabilities
- Time intelligence

Multidimensional views

Multidimensional views are inherently representative of an actual business model. Rarely is a business model limited to fewer than three dimensions. Managers typically look at financial data by scenario (for example, actual versus budget), organization, line items, and time; and at sales data by product, geography, channel, and time.

A multidimensional view of data provides more than the ability to “slice and dice”; it provides the foundation for analytical processing through flexible access to information. Database design should not prejudice which operations can be performed on a dimension or how rapidly those operations are performed. Managers must be able to analyze data across any dimension, at any level of aggregation, with equal functionality and ease. OLAP software should support these views of data in a natural and responsive fashion, insulating users of the information from complex query syntax. After all, managers should not have to understand complex table layouts, elaborate table joins, and summary tables.

Whether a request is for the weekly sales of a product across all geographical areas or the year-to-date sales in a city across all products, an OLAP system must have consistent response times. Managers should not be penalized for the complexity of their queries in either the effort required to form a query or the amount of time required to receive an answer.

Calculation-intensive capabilities

The real test of an OLAP database is its ability to perform complex calculations. OLAP databases must be able to do more than simple aggregation. While aggregation along a hierarchy is important, there is more to analysis than simple data roll-ups. Examples of more complex calculations include share calculations (percentage of total) and allocations (which use hierarchies from a top-down perspective).

Key performance indicators often require involved algebraic equations. Sales forecasting uses trend algorithms, such as moving averages and percentage growth. Analyzing the sales and promotions of a given company and its competitors requires modeling complex relationships among the players. The real world is complicated — the ability to model complex relationships is key in analytical processing applications.

Time intelligence

Time is an integral component of almost any analytical application. Time is a unique dimension, because it is sequential in character (January always comes before February). True OLAP systems understand the sequential nature of time. Business performance is almost always judged over time, for example, this month versus last month, this month versus the same month last year.

The time hierarchy is not always used in the same manner as other hierarchies. For example, a manager may ask to see the sales for May or the sales for the first five months of 1995. The same manager may also ask to see the sales for blue shirts but would never ask to see the sales for the first five shirts. Concepts such as year-to-date and period over period comparisons must be easily defined in an OLAP system.

In addition, OLAP systems must understand the concept of balances over time. For example, if a company sold 10 shirts in January, five shirts in February, and 10 shirts in March, then the total balance sold for the quarter would be 25 shirts. If, on the other hand, a company had a head count of 10 employees in January, only five employees in February, and 10 employees again in March, what was the company's employee head count for the quarter? Most companies would use an average balance. In the case of cash, most companies use an ending balance.

2.3.3 Statistics

Statistical tools are typically used to address the business problem of generating an overview of the data in your database. This is done by using techniques that summarize information about the data into statistical measures that can be interpreted without requiring every record in the database to be understood in detail (for example, the application of statistical functions like finding the maximum or minimum, the mean, or the variance). The interpretation of the derived measures require a certain level of statistical knowledge.

These are typical business questions addressed by statistics:

- What is a high-level summary of the data that gives me some idea of what is contained in my database?
- Are their apparent dependencies between variables and records in my database?
- What is the probability that an event will occur?
- Which patterns in the data are significant?

To answer these questions the following statistical methods are typically used:

- Correlation analysis
- Factor analysis

- Regression analysis

These functions are detailed in 7.2.2, “Statistical functions” on page 175.

2.3.4 Data mining

However, in contrast with statistical analysis, data mining analyzes all the relevant data in your database and extracts hidden patterns.

Data mining is to some extent based on the techniques and disciplines used in statistical analysis. However, the algorithms used in data mining automate many of the tedious procedures that you would need to go through to obtain the same depth of analysis using traditional statistical analysis.

An introduction to data mining is given in Chapter 3, “A generic data mining method” on page 23.

2.4 Data warehouse, OLAP and data mining summary

If the recommended method when building data warehouses and OLAP datamarts is:

- To build an ODS where you collect and cleanse data from OLTP systems.
- To build a star schema data warehouse with fact table and dimensions tables.
- To use data in the data warehouse to build an OLAP datamart.

Then, the recommended method for building data warehouses and data mining datamarts could be quite the same:

- To build an ODS where you collect and cleanse data from OLTP systems.
- To build a star schema data warehouse with fact table and dimensions tables.
- To pick the dimension which is of main interest, for example, customers — to use aggregation and pivot on the fact table and maybe one or two other dimensions in order to build a flat record schema or a datamart for the mining techniques.

As a star schema model or multidimensional model, a data warehouse should be a prerequisite for OLAP datamarts, even if it is not a prerequisite for a data mining project, it may help as a design guideline.

OLAP and data mining projects could use the same infrastructure. The construction of the star schema and extracting/transforming/loading steps to build the data warehouse are the responsibilities of the IT department. An IT department should of course take into account the business users' requirements on OLAP as cubes or multidimensional databases, reports, and also data mining models to design the data warehouse.

OLAP and data mining can use the same data, the same concepts, the same metadata and also the same tools, perform in synergy, and benefit from each other by integrating their results in the data warehouse.

3

A generic data mining method

Data mining is one of the main applications that are available to you as part of your overall BI architecture. You may already use a number of analysis and reporting tools to provide you with the day to day information you need. So why is data mining different from the normal types of statistical analysis and other business reporting tools that you use?

In this chapter we describe what data mining is all about and describe some of the things that you can do with the tools and techniques that data mining provides. Gaining an understanding of what data mining can do will help you to see the types of business questions that you can address and how you can take the first steps along the road of mining your own business. To help in this respect we have developed a generic data mining method that you can use as a basic guide. The generic method is explained and in the following chapters we will show how it can be applied to address specific retail business issues.

3.1 What is data mining?

Data mining is treated by many people as more of a philosophy, or a subgroup of mathematics, rather than a practical solution to business problems. You can see this by the variety of definitions that are used, for example:

> *"Data mining is the exploration and analysis of very large data with automatically or semi-automatically procedures for previously unknown, interesting, and comprehensible dependencies"*
>
> Or
>
> *"Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."*

While these definitions have their place, in this book we want to concentrate on the practical issues of data mining and show how to make data mining work for your retail business. Specifically, we want to show you what you have to do to successfully mine your own business and to end up with reliable results that you can put to use.

Although data mining as a subject in its own right, it has only existed for less than 10 years, and its origins can be traced to the early developments in artificial intelligence in the 1950's. During this period, developments in pattern recognition and rule based reasoning were providing the fundamental building blocks on which data mining was to be based. Since this time, although they were not given the title of data mining, many of the techniques that we use today have been in continuous use, primarily for scientific applications.

With the advent of the relational database and the capability for commercial organizations to capture and store larger and larger volumes of data, it was released that a number of the techniques that were being used for scientific applications could be applied in a commercial environment and that business benefits could be derived. The term data mining was coined as a phrase to encompass these different techniques when applied to very large volumes of data. Figure 3-1 shows the developments that have taken place over the past 40 years.

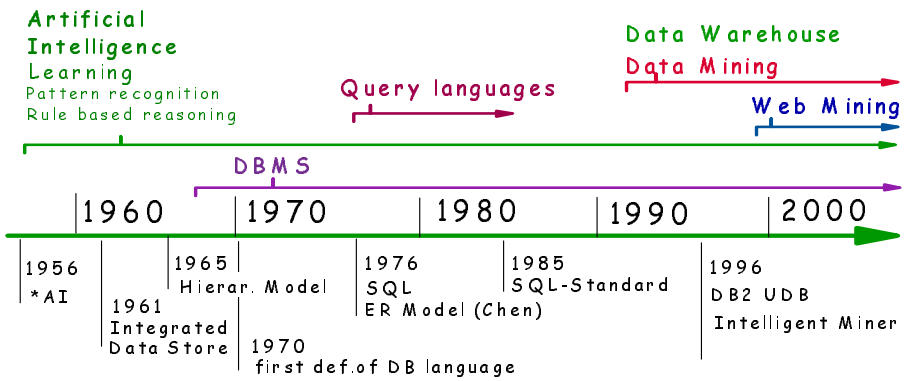


Figure 3-1 A historical view of data mining

Some of the techniques that are used to perform data mining are computationally complex, and in order to discover the patterns existing within large data sets they have to perform a large number of computations. In the last 10 years the growth in the use of large commercial databases (specifically data warehouse) coupled with the need to understand and interpret this data and the availability of relatively inexpensive computers has lead to an explosion in the use of data mining for a wide variety of commercial applications.

3.2 What is new with data mining?

Data mining is about discovering new things about your business from the data you have collected. You might think that you already do this using standard statistical techniques to explore your database. In reality what you are normally doing is making a hypothesis about the business issue that you are addressing and then attempting to prove or disprove your hypothesis by looking for data to support or contradict the hypothesis.

For example, suppose that as a retailer, you believe that customers from “out of town” visit your larger inner city stores less often than other customers, but when they do so they make larger purchases. To answer this type of question you can simply formulate a database query looking, for example, at your branches, their locations, sales figures, customers and then compile the necessary information (average spend per visit for each customer) to prove your hypotheses. However, the answer discovered may only be true for a small highly profitable group of out-of-town shoppers who visited inner-city stores at the weekend. At the same

time, out-of-town customers (perhaps commuters) may visit the store during the week and spend exactly the same way as your other customers. In this case, your initial hypothesis test may indicate that there is no difference between out-of-town and inner-city shoppers.

Data mining uses an alternative approach beginning with the premise that you do not know what patterns of customer behaviors exist. In this case you might simply ask the question, what are their relationships (we sometimes use the term *correlations*) between what my customers spend and where they come from? In this case, you would leave it up to the data mining algorithm to tell you about all of the different types of customers that you had. This should include the out-of-town, weekend shopper. Data mining therefore provides answers, without you having to ask specific questions.

The difference between the two approaches is summarized in Figure 3-2.

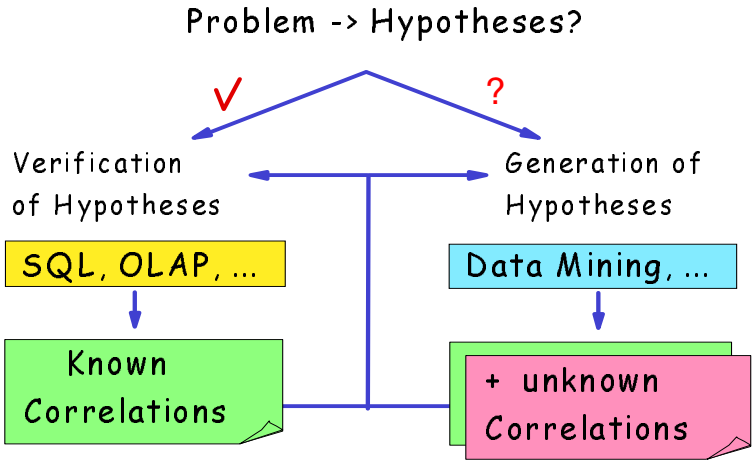


Figure 3-2 Standard and data mining approach on information detection

So how do you set about getting the answers to the sorts of business issues that data mining can address. This is usually a complex issue, but that is why we have written this book. To help in this regard, we follow a **generic method** that can be applied to a wide range of business questions, and in the following chapters we show how it can be applied to solve chosen business issues.

3.3 Data mining techniques

As we explained previously, a variety of techniques have been developed over the years to explore for and extract information from large data sets. When the name data mining was coined, many of these techniques were simply grouped together under this general heading and this has led to some confusion about what data mining is all about. In this section we try to clarify some of the confusion.

3.3.1 Types of techniques

In general, data mining techniques can be divided into two broad categories:

- Discovery data mining
- Predictive data mining

Discovery data mining

Discovery data mining is applied to a range of techniques which find patterns inside your data without any prior knowledge of what patterns exist. The following are examples of discovery mining techniques.

Clustering

Clustering is the term for a range of techniques which attempts to group data records on the basis of how similar they are. A data record may, for example, comprise a description of each of your customers. In this case clustering would group similar customers together, while at the same time maximizing the differences between the different customer groups formed in this way. As we will see in the examples described in this book, there are a number of different clustering techniques, and each technique has its own approach to discovering the clusters that exist in your data.

Link analysis

Link analysis describes a family of techniques that determines associations between data records. The most well known type of link analysis is market basket analysis. In this case the data records are the items purchased by a customer during the same transaction and because the technique is derived from the analysis of supermarket data, these are designated as being in the same basket. Market basket analysis discovers the combinations of items that are purchased by different customers, and by association (or linkage) you can build up a picture of which types of product are purchased together. Link analysis is not restricted to market basket analysis. If you think of the market basket as a grouping of data records then the technique can be used in any situation where there are a large number of groups of data records.

Frequency analysis

Frequency analysis comprises those data mining techniques that are applied to the analysis of time ordered data records or indeed any data set that can be considered to be ordered. These data mining techniques attempt to detect similar sequences or subsequences in the ordered data.

Predictive Mining

Predictive data mining is applied to a range of techniques that find relationships between a specific variable (called the *target variable*) and the other variables in your data. The following are examples of predictive mining techniques.

Classification

Classification is about assigning data records into pre-defined categories. For example, assigning customers to market segments. In this case the target variable is the category and the techniques discover the relationship between the other variables and the category. When a new record is to be classified, the technique determines the category and the probability that the record belongs to the category. Classification techniques include decision trees, neural and Radial Basis Functions (RBF) classifiers.

Value prediction

Value prediction is about predicting the value of a continuous variable from the other variables in a data record. For example, predicting the likely expenditure of a customer from their age, gender and income group. The most familiar value prediction techniques include linear and polynomial regression, and data mining extends these to other techniques, such as neural and RBF value prediction.

3.3.2 Different applications that data mining can be used for

There are many types of applications to which data mining can be applied. In general, other than for the simplest applications, it is usual to combine the different mining techniques to address particular business issues. In Figure 3-3 we illustrate some of the types of applications, drawn from a range of industry sectors, where data mining has been used in this way. These applications range from customer segmentation and market basket analysis in retail, to risk analysis and fraud detection in banking and financial applications.

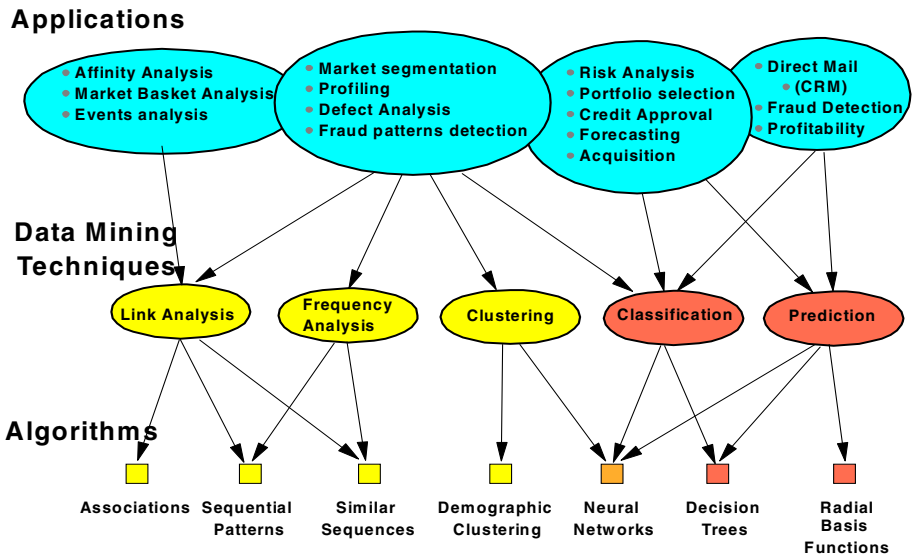


Figure 3-3 From applications to algorithms

Because there is such a bewildering range of things that can be done using data mining, our objective is to concentrate on some of the techniques that apply to your specific industry sector. To assist you in the process of understanding how to decide what techniques are applicable and how to set about the whole process of mining you own business, we have developed a generic data mining method. The objective is to define a sequence of steps that can be followed for all data mining activities, and in subsequent chapters we show how the method is applied to address specific business issues.

3.4 The generic data mining method

This section describes the generic data mining method which comprises seven steps; these are:

- Defining the business issue in a precise statement
- Defining the data model and the data requirements
- Sourcing data from all available repositories and preparing the data (The data could be relational or in flat files, stored in a data warehouse, computed, created on-site or bought from another party. They should be selected and filtered from redundant information.)
- Evaluating the data quality

- Choosing the mining function and defining the mining run
- Interpreting the results and detecting new information
- Deploying the results and the new knowledge into your business

These steps are illustrated in Figure 3-4 and the following sections expand on each of the stages.

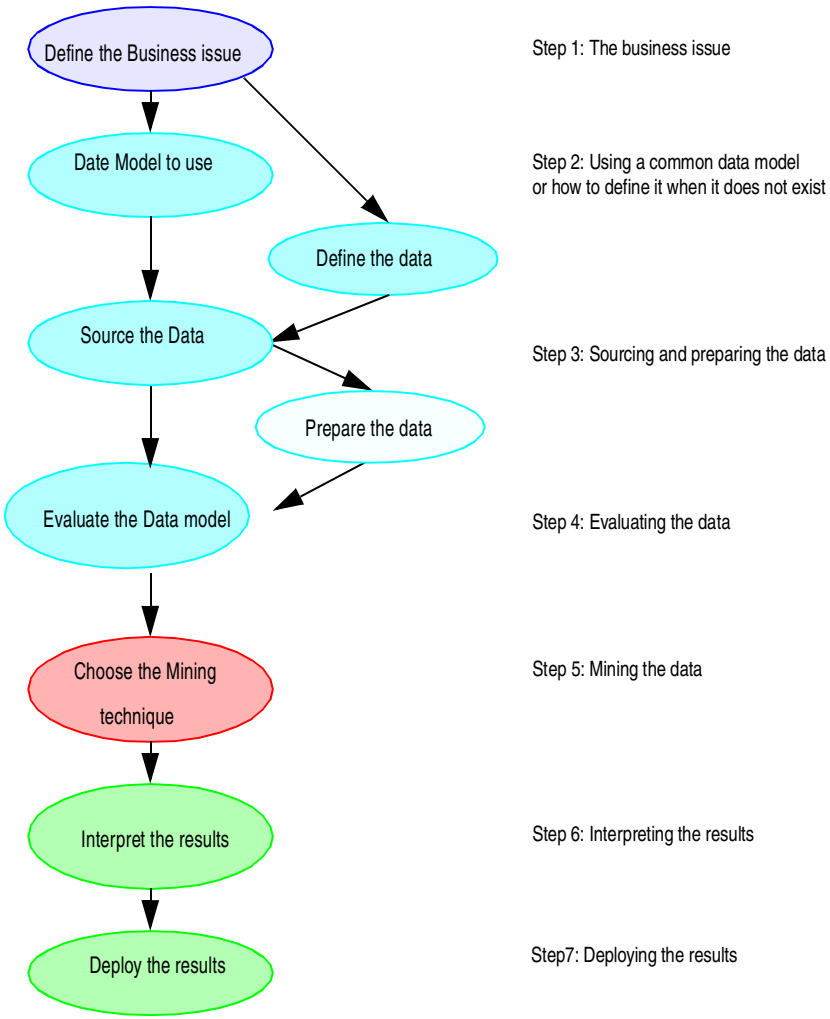


Figure 3-4 The generic method

3.4.1 Step 1 — Defining the business issue

All to often, organizations approach data mining from the perspective that there must be some value in the data we have collected, so we will just use data mining and discover what's there. Using the mining analogy, this is rather like choosing a spot at random and starting to dig for gold. This may be a good strategy if the gold is in large nuggets, and you were lucky enough to choose the right spot to dig, but if the gold could only be extracted by "panning" or some other technique, you may spend a lot of time throwing away the valuable material in a fruitless search, for searching for the right thing at the wrong place or with the wrong technique.

Data mining is about choosing the right tools for the job and then using them skillfully to discover the information in your data. We have already seen there are a number of tools that can be used, and that very often we have to use a combination of the tools at our disposal, if we are to make real discoveries and extract the value from our data.

The *first step in our data mining method* is therefore to identify the business issue that you want to address and then determine how the business issue can be translated into a question, or set of questions, that data mining can address.

By *business issue* we mean that there is an identified problem to which you need an answer, where you suspect, or know, that the answer is buried somewhere in the data, but you are not sure where it is.

A business issue should fulfill the requirements of having:

- A clear description of the problem to be addressed
- An understanding of the data that might be relevant
- A vision for how you are going use the mining results in your business

Describing the problem

If you are not sure what questions data mining can address, then the best approach is to look at examples of where it has been successfully used, either in your own industry or in related industries. Many business and research fields have been proven to be excellent candidates for data mining. The major fraction are covered by banking, insurance, retail and telecommunications (telecoms), but there are many others such as manufacturing, pharmaceuticals, biotechnology and so on, where significant benefits have also been derived. Well-known approaches are: customer profiling and cross-selling in retail, loan delinquency and fraud detection in banking and finance, customer retention (attrition and churn) in telecoms, patient profiling and weight rating for Diagnosis Related Groups in health care and so on. Some of these are depicted in Figure 3-5.

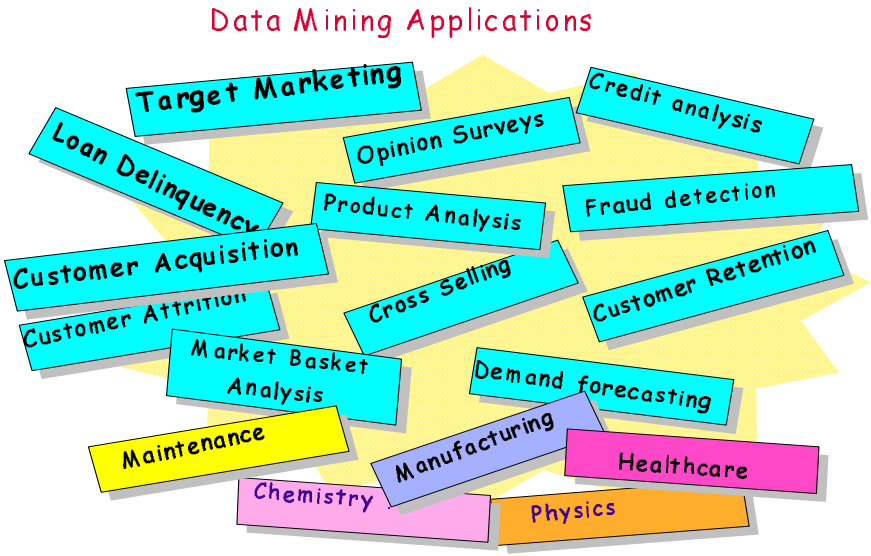


Figure 3-5 Business and research applications

The objective behind this book and others in the series is to describe some of these different issues and show how data mining can be used to address them.

Even where the specific business issue you are trying to address has not been addressed elsewhere, understanding how data mining can be applied will help to define your issue in terms that data mining can answer. You need to remember that data mining is about the discovery of patterns and relationships in your data. All of the different applications are using the same data mining concepts and applying them in subtly different ways.

With this in mind, when you come to define the business issue, you should think about it in terms of patterns and relationships. Take fraud as an example. Rather than ask the question can we detect fraudulent customers, you could ask the question, can we detect a small group of customers who exhibit unusual characteristics that may be indicative of fraud? Alternatively, if you have identified some customers who are behaving fraudulently, the question is, can you identify some unique characteristics of these customers that would enable you to identify other potentially fraudulent customers?

Understanding your data

As you are formulating the business question, you need to also think about whether the data that you have available is going to be sufficient to answer the question. It is important to recognize that the data you hold may not contain the information required to enable you to answer the question you are posing. For example, we suppose you are trying to determine why you are losing customers and the reason is that your competitors are undercutting you on price. If you do not have competitor pricing data in your database, then clearly data mining is not going to provide the answer. Although this is a trivial example, sometimes it is not so obvious that the data cannot provide the answer you are looking for. The amazing thing is how many people still believe that data mining should be able to perform the impossible.

Where the specific business issue has been addressed elsewhere, then knowing what data was used to address the issue will help you to decide which of your own data should be used and how it may need to be transformed before it can be effectively mined. This process is termed the construction of a common data model. The use of common data models is a very powerful aid to performing data mining as we show when we address specific business issues.

Using the results

When defining the business issue that you want to address with data mining, it is important that you think carefully about how you are going to use the information that you discover. Very often, by considering how you are going to deploy the results of your data mining into your business, will help to clarify the business issue you are trying to address and determine what data you are going to use.

Suppose for example, that you want to use data mining to identify which types of existing customers will respond to new offers or services and then use the results to target new customers. Clearly the variables you use when doing the data mining on your existing customers, must be the same variables that you can derive about your new customers. In this case you cannot use the 6-month aggregated expenditure (*aggregated spend*) on particular products if all you have available for new customers is the expenditure from a single transaction. Thinking about how you are going to use the information you derive places constraints on the selection of data that you can use to perform the data mining and is therefore a key element in the overall process of translating the business issue into a data mining question.

3.4.2 Step 2 — Defining a data model to use

The *second step in the generic data mining method* is to define the data to be used. A business like yours can collect and store vast amounts of data. Usually the data is being stored to support various applications and as we considered in 2.1, “Business Intelligence” on page 8; the best way to do this is to use some form of data warehouse. Although not all data warehouse architectures are the same, one way they can be used efficiently to support your applications is shown in Figure 3-6. In this case each end user application is supported by its own datamart which is updated at regular intervals or when specific data changes, to reflect the needs of the application.

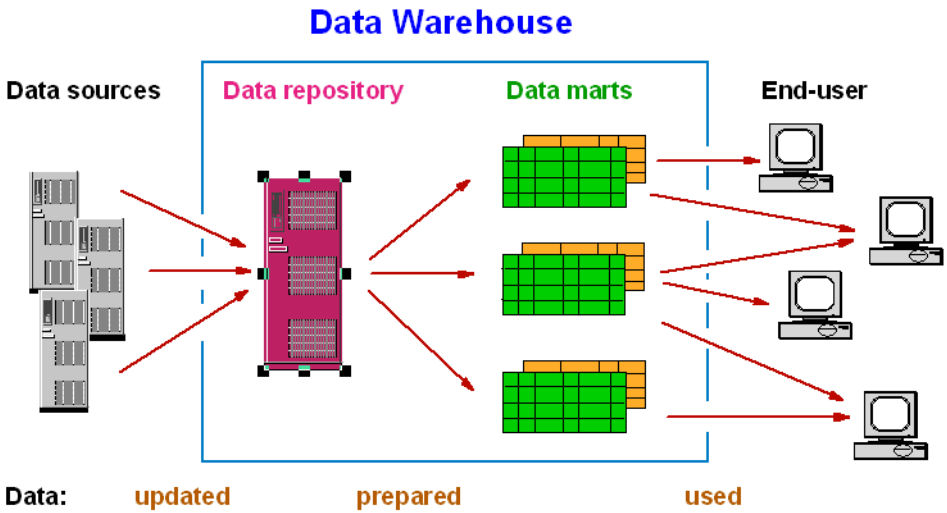


Figure 3-6 Data warehouse architecture

In this structure each datamart has its own specific data and holds knowledge about how the data was derived, the data format used, what aggregations have been performed, what data cleansing has been done and so on. In 2.2.6, “Metadata” on page 12, we described this additional information as metadata. Where the data is being used routinely to support a specific business application, the data and metadata together form what we call a *data model* that supports the application.

Typically the data model will define:

- Data sources used
- Data types
- Data content
- Data description

- Data usage

The *data sources* indicate the physical location from where the data is derived or stored. The *data type* defines how the data is structured (for example, the date time format used). The *data content* lists the tables or data files and the fields which they contain. The *data description* delivers the names and description of these fields. The *data usage* considers the ownership of tables and fields, how users understand their content, and, although often neglected, how the users exploit them. The data model also contains information on when the data is valid, when it must be replicated and so on.

Data mining is just another application and, depending on what it is being used for, requires its own data model. For most data mining applications, the data model required is in the form of a single file or database table, with one record per customer, or department, or whatever is the target of the investigation.

Note: In database terms, the required table is called a denormalized table and can be either a physical table in the database or a database “view” of joined tables, each comprising some of the variables consisting of one or more variables.

Each record can comprise one or many variables, where each variable may be derived from a number of different data sources but are tied to the same target variable (for example, the customer) in some way. In most business applications the most common data types are:

- Transactional data
- Relationship data
- Demographic data

Transactional data is operational data generated each time some interaction occurs with the target. This data typically contains a timestamp and some transaction identifier together with details of the transaction. This type of data may, for example, relate to point of sales data for a customer in a supermarket, or to the recording of information on a production line in a manufacturing application.

Relationship data is the nonvolatile data containing relatively stable information about customers, products, equipment, items, and working processes.

Demographic data comprises person-specific (customer, patient) data usually coming from external sources. Typically this includes information on age, gender postal code and so on.

Use of common data models

Defining data models for any application is often a complex task and defining data models for data mining is no exception. Where the data model is required to support an application that has specific requirements (for example, some form of business reporting tool) then the data can be defined by asking the end users what types of information they require and then performing the necessary aggregations to support this requirement. In the case of data mining, the challenge is that very often you are not sure at the outset which variables are important and therefore exactly what is required. Generating data models for completely new data mining applications can therefore be a time consuming activity.

The alternative is to use common data models that have been developed to solve similar business issues to the ones you are trying to address. While these types of models may not initially provide you with all of the information you require, they are usually designed to be extendable to include additional variables. The main advantage of using a common data model is that it will provide you with a way of quickly seeing how data mining can be used within your business. In the following chapters we suggest some simple data models that can be used in this way.

3.4.3 Step 3 — Sourcing and preprocessing the data

The *third step in the generic data mining method* is the sourcing and preprocessing of the data that populates the data model. Having a defined data model provides the necessary structure, in terms of the variables that we are going to mine, but we still have to provide the data.

Data sourcing and preprocessing comprises the stages of *identifying, collecting, filtering* and *aggregating* (raw) data into a format required by the data models and the selected mining function. Since sourcing and preparing the data are the most time consuming parts of any data mining project, we describe these crucial steps in broader detail. Where the data is derived from a data warehouse, many of these stages will already have been performed.

The data sources

The data sources can be different by origin and content as shown in Figure 3-7.

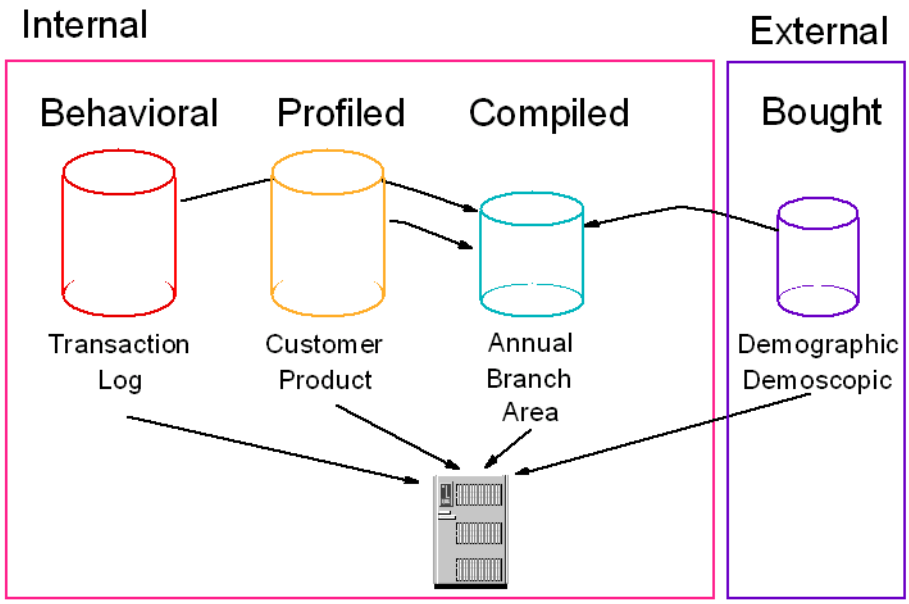


Figure 3-7 Data sources by origin and content

Every business uses standard internal data sources. Many of them are similar from their point of content. Therefore, a *customer database* or a *product database* could be found in nearly any data scenario.

Data mining, in common with many other analysis tools, usually requires the data to be in one consolidated table or file. If the variables required are distributed across a number of sources then this consolidation must be performed such that a consistent set of data records is produced. If the data is stored in relational database tables then the creation of a new tables or a database view is relatively straight forward, although where complex aggregations are required this can have a significant impact on database resources.

Data preprocessing

If the data has not been derived from a data warehouse then the data preprocessing functions of cleansing, aggregated, transforming, and filtering, that we described in 2.2, “Data warehouse” on page 8, must be undertaken. Even when the data is derived from a data warehouse, there may still be some additional transformations of the data that need to be performed before mining can proceed. Structurally the *data preprocessing* can be displayed as in Figure 3-8.

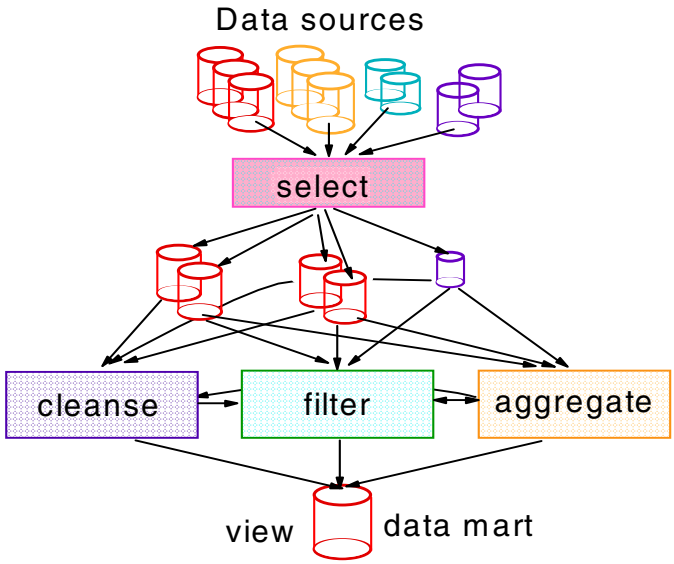


Figure 3-8 Data preprocessing

Data mining tools usually provide limited capability to cleanse the data, because this is a specialized process and there are a number of products that can be used to do this efficiently. Aggregation and filtering can be performed in a number of different ways depending on the precise structure of your data sources. Some of the tools available to do this with the IM for Data product are described in 7.2.1, “Data preparation functions” on page 173.

3.4.4 Step 4 — Evaluating the data model

Having populated the data model with data we still have ensure that the data used to populate our data model fulfills the requirement of completeness, exactness and relevance. To do this we perform the *fourth step in the generic data mining method,* which is to perform an initial evaluation; the steps are described in Figure 3-9.

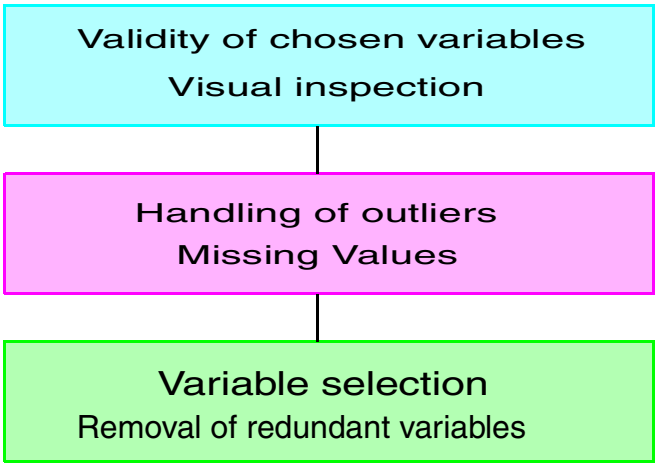


Figure 3-9 Overview: Steps of data evaluation

The first step *visual inspection* comprises browsing the input data with visualizing tools. This may lead to the detection of *implausible distributions*. For example, a wrong join of tables during the data preparation step could result in variables containing values actually belonging to different fields.

The second step deals with the *identification of inconsistencies* and the *resolution of errors*. Unusual distributions found within the first step could be induced by a badly done data collection. Many *outlying and missing values* produce biased results. For example, the interesting information of a correlation between variables indicating the behavior of a customer group and variables describing new offerings could be completely suppressed if too many missing values are accepted.The mitigation of outliers and the transformation of missing values in meaningful information however improves the data quality enormously. Many data mining functions take account of a minor fraction of missing values. But an early treatment of missing values can prevent us from biased mining results and unconsciously progressed errors.

The last step is the final *selection of features/variables* for the mining run. Variables could be superfluous by presenting the same or very similar information as others, but increasing the run time. *Dependent or highly correlated variables* could be found with statistical tests like bivariate statistics, linear and polynomial regression. Dependent variables should be reduced by selecting one variable for all others or by composing a new variable for all correlated ones by factor or component analysis.

Not all variables remaining after the statistical check are nominated as input; only variables with a clear interpretation and variables that make sense for the end user should be selected. A proven data model simplifies this step. The selection of variables in that stage can indeed only be undertaken with practical experience in the respective business or research area.

3.4.5 Step 5 — Choosing the data mining technique

Besides the steps defining business issues, data modeling, and preparation, data mining also comprises the crucial step of the selection of the best suited mining technique for a given business issue. This is the *fifth step in the generic data mining method*. This step not only includes defining the appropriate technique or mix of techniques to use, but also the way in which the techniques should be applied.

Several types of techniques (or algorithms) are available:

- Classification
- Associations
- Clustering
- Value prediction
- Similar patterns
- Similar time sequences

Others have been developed for the detection of different kinds of correlations and patterns inside databases.

The selection of the method to use will often be obvious, for example, market basket analysis in retail will use the associations technique which was originally developed for this purpose. However, the associations technique could be applied to a diverse range of applications, for example, to discover the relationships between faults occurring in production runs and the sources from which the components were derived.

The challenge is usually not which technique to use but the way in which the technique should applied. Because all of the data mining techniques require some form of parameter selection, this then requires experience of how the techniques work and what the various parameters do. In the examples given in the following chapters, we describe some of the parameters that need to be defined and how this can be done.

3.4.6 Step 6 — Interpreting the results

Interpreting the results is *the sixth step in the generic mining method*. The results from performing any type of data mining can provide a wealth of information that can sometimes be difficult to interpret. Our experience is that the interpretation phase requires the input from a business expert who can translate the mining results back into the business context. Because we do not expect the business analyst to necessarily be a data mining specialist, it is important that the results are presented in such a way that they are relatively easy to interpret.

To assist in the interpretation process, it is necessary to have at your disposal a range of tools that enable you to visualize the results and to provide the necessary statistical information that you need to make the interpretation.

In the following chapters, we provide you with a number of examples of the types of visualization techniques that are available and how to understand what the different results are telling you about your business.

3.4.7 Step 7 — Deploying the results

The *seventh and final step in the generic data mining* method is perhaps the most important of all. It deals with the question of how to deploy the results of the data mining into your business. If you only see data mining as an analytical tool, then you are failing to realize the full potential of what data mining has to offer.

As we have already explained, when you perform data mining you can both discover new things about your customers and determine how to classify or how to predict particular characteristics or attributes. In all these cases data mining creates mathematical representations of the data that we call models. These models are very important, because they not only provide you with a deeper insight of your business but can themselves be deployed in or used by other business processes, for example, your CRM systems.

When embarking on any data mining activity you should think carefully about the way in which you intend to use the data mining results and where in your business the results will have the greatest impact. In the following chapters we describe some of the ways in which both data mining results and data mining models can be used.

One particular important development in regard to the deployment of the data mining results is the development of standards for exchanging data mining models and of being able to deploy these models directly into relational databases, such as DB2 Universal Database and ORACLE. The new standard is

based on what is called the Predictive Model Markup Language (PMML). This standard provides for the exchange of analytic models like linear regression, clustering, decision tree, and neural network. The most important advantages are:

- Data miner experts on-site are not necessary
- Computational resources are exploited efficiently
- It allows real time (event-triggered) processing and batch processing
- It enables foreign software applications access to the modeling logic
- It allows the generalization of any future model type

Further details on DB2 IM Scoring are presented in 7.3, “DB2 Intelligent Miner Scoring” on page 179.

3.4.8 Skills required

To successfully implement a data mining project using the above method, you will require a mix of skills in the following areas:

- Data manipulation (for example SQL)
- Knowledge of mining techniques
- Domain knowledge or ability to communicate with domain experts
- Creativity

These skills are normally not being incorporated in one person and Figure 3-10 shows the structure of a typical data mining team.

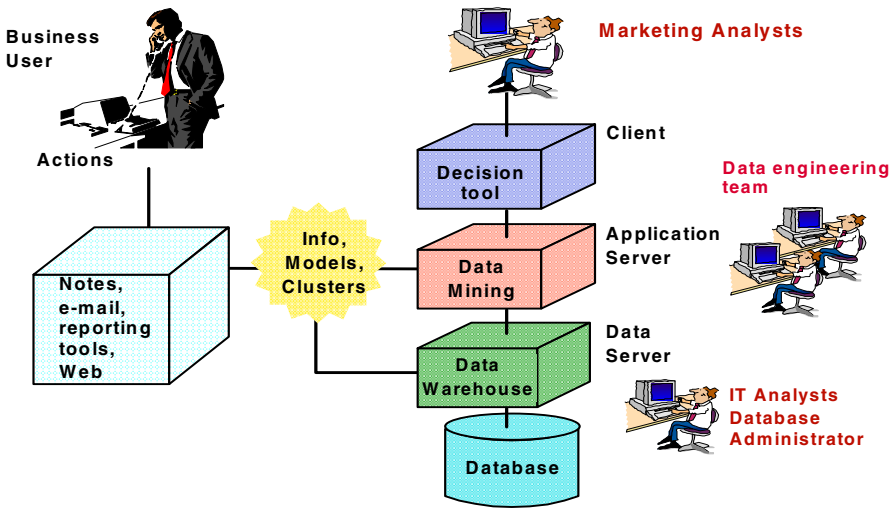


Figure 3-10 The data mining team

Such a team will comprise the following:

- *Marketing analyst* who is informed in the branches of businesses which have been selected for data mining.
- *IT analyst* is needed who is experienced with data management procedures within your business.
- *Data engineering team* who will have the lead and the experience in all data mining topics.
- *Business user* should be selected who can check the usability of the mining result and evaluate the deployment from a solely business perspective. It should be mentioned that not few data mining projects run into problems by underestimating the efforts of searching and delivering raw data in a reasonable quality.
- *Project owner* who is normally the head of a branch inside the company, should support the work and help to resolve problems.

Whether or not these are different individuals clearly depends on the mix of skills that they have, but in general the team must be able to accomplish the following:

- Understanding the data source: There are two aspects of understanding the data source, which are the knowledge about the physical data situation in the company and the usage of the data proposed for data mining. Normally, the data mining expert is not the data administrator, who is responsible for the all data repositories. In this case the interaction with the database owner and the mining expert must be guaranteed.
- Understanding the data preparation: Data preparation needs a lot of expertise in creating new data input (for example, SQL programming skill) and a good understanding of the given raw data and their content. An excellent data miner may not be successful if he/she lacks expertise in the business field under discussion.
- Understanding the algorithms: Using algorithms means to be able to adapt the setting for the various mining runs. Because all data mining functions are highly sophisticated from a implementation point of view, data mining experts are demanded who are well trained with the selected data mining toolkit. Namely, these persons must overview how much effort has to be undertaken to solve single steps of the generic methods (Figure 3-11), and how to solve each task either with the toolkit, or with the database stem, or with additional statistical functions.
- Understanding the business needs: This concerns the interaction with the business end users.

3.4.9 Effort required

It is difficult to be prescriptive about the amount of effort that will be required to perform a typical data mining project. If you were starting from a position of having no data warehouse and with data in disparate files and databases then the type of effort profile required is shown in Figure 3-11.

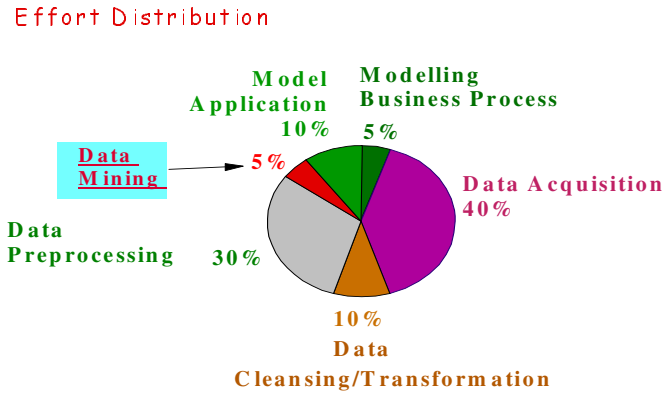


Figure 3-11 Effort distribution

You can see immediately that the vast majority of the effort is spent in the data preparation activities. If you already have the data in a usable form, then the task of mining your own business becomes much easier.

4

How can I characterize my customers from the mix of products that they purchase?

In an ideal world all retail organization's would like to be in a position to treat each customer as an individual and to maximize customer satisfaction and profitability on a one-to-one basis. This is the ultimate goal of Customer Relationship Management (CRM). In the real world, one-to-one interaction is in general not economically viable and to address these challenges most retail organizations attempt to segment their customers into a relatively small number of market segments.

By using data mining you can derive customer segments directly from the data, as opposed to using pre-defined business definitions. The data driven approach provides a natural grouping of your customers based on the customer attributes that you can measure directly or derive indirectly from other sources of information.

In this chapter we will describe some of the data mining methods that can be used for discovering these data derived segments and how you can use them to characterize your customers and map out your business.

4.1 The business issue

Understanding your customers is key to any successful business, however when you are in a highly dynamic business, such as retailing, with large numbers of customers and continuously changing products and services, this becomes a significant challenge. What are my customers doing? How are they reacting to the different initiatives that I am taking? How can I develop and offer new products and services? These are just some of the questions that confront you.

Customers are also increasingly coming to expect and demand higher levels of personal service. While in an ideal world you would like to treat each customer as an individual, even if you had all of the relevant information and the means to deliver this level of service, trying to determine what is appropriate seems virtually impossible. The natural solution to the problem is to attempt to identify groups of customers that exhibit similar characteristics, for example, the types of products they purchase. Grouping or segmenting customers in this way enables you to develop sales and marketing strategies that are targeted on the group rather than the individual. At the same time gives each customer within the group the impression that you are addressing them personally. Campaign management systems (for example, Xchange Campaign) precisely use this approach.

Using customer segmentation to drive your CRM, the main challenge then becomes that of how to group your customers in an appropriate way. This grouping is usually performed using some type of business rule. Such rules might seek to group customers by their total spend and frequency of visit, or by some particular brand of product that they purchase. There are many variations in the types of business rule that can be developed.

In general it is difficult to create rules that describe more than a few customer attributes (for example, frequency of visit, total expenditure, types of products purchased). In each of these attributes customers are typically ranked into high, medium and low categories and these are given grand sounding names like “Frequent Visitor”, “Core Customer” or “Hit and Run”. A sample business rule may be:

> *If the customer is a “High Spender” and a “Frequent Visitor” and purchases “Store Own Brand” products, then they are “Loyal” customers.*

These rules are relatively simple, because it becomes increasingly difficult to use more than two or three customer attributes at any one time. In fact most retail, and indeed many other types of organization, end up with business rules that combine either three customer attributes with high or low ranking, or two characteristics in with high, medium and low ranking.

Note: It is important to realize that business rules must in some way ensure that the rules are exclusive, in the sense that customers are only categorized by one rule and allocated to one segment. So if more attributes are used, or there are more levels of ranking, it becomes increasingly difficult to keep all of the combinations in your head and so some attributes are inevitably ignored in some rules and included in others.

Although the end results look impressive, you nearly always end up with approximately eight or nine customer groups that are then grandly identified as *market segments* and promotional campaigns are then developed to address these segments. You may be paying consultants large amounts of money to produce these segments for you, but essentially this is what you are getting for your money.

This approach however raises a number of questions: Are these the best segments that I could produce? Did I choose the right characteristics? Am I missing some other combination of characteristics that could be important? How can data mining help me do a better job?

4.1.1 How can data mining help?

In data mining we also use the term segmentation, although the data mining technique that we use to perform segmentation is clustering. The objective of the segmentation produced by data mining is to discover the different groups of customers that are suggested using the data you hold about your customers, rather than by you having to make some judgement about what are the most important characteristics. In other words, it suggests to you what the business rules should be, rather than what you think they might be.

Note: *Clustering* means that we try to group customers together who have similar characteristics, while at the same time trying to maximize the difference between customers in the different groups that we create.

Data mining will allow you to identify customer types that you did not recognize you had, and will inevitably lead to new ideas about your market segments and the most appropriate way to offer new products and services to the new customer groups that you discover. So what do you have to do to achieve this? In 3.4, “The generic data mining method” on page 29, we outlined a generic mining method and in the following sections we describe how this method can be applied to discover the customer segments in your retail business and how these can be used in your CRM systems.

4.1.2 Where should I start?

The first stage in the generic method is to translate the business issue you are trying to address, into a question, or set of questions that can be addressed by data mining. In the case of customer segmentation this is relatively straight forward, since the data mining technique of clustering does precisely what you are trying to achieve. However, we have to recognize that we are now talking about a data driven segmentation and therefore we have to consider exactly what data is required.

To address this question you have to consider two things:

- How am I going to interpret the results of the segmentation?
- How am I going to use the segmentation results?

If you are going to produce a data driven segmentation, based on some attributes that you can measure or derive about your customers, then the result of performing the data driven segmentation will be expressed in terms of these same attributes. Because data mining can use many attributes to create the segmentation, it is tempting to just generate as many attributes as you can think of and let the data mining use whatever attributes are necessary to generate the segmentation. While the data mining tools that we use can do this, the results that you get can become very difficult to interpret and in general this is not a good approach to take. In the same way, you also need to think about how the mining results are going to be used within your business and how the customer attributes you are using to perform the segmentation can be generated and incorporated into existing CRM systems.

In general a good place to start with data mining is to look at your existing CRM processes and think carefully about what exactly you are trying to achieve through the data mining that you are going to perform. A good example of this would be where you have an existing set of business rules that you use to segment your customers. Although we were a little disparaging about business rules, they do provide a view of your business that you understand and work with. For example, such rules may already form an integral part of your CRM system or may be the basis for much of the strategic planning within your business. Our experience is that by beginning the data mining using these rules, you are able to compare the results that you get with your existing customer segmentation. Data mining will enable you and your colleagues to immediately understand how the data mining results can be interpreted and applied. It also enables you to see if the segmentation suggested by data mining conforms with the market segments you currently have. Usually this is not the case, but what is often suggested is an interesting variation on what you already have, and it is then much easier to make decisions about how to modify your current business process in response. The way in which you can perform this type of data mining segmentation is described in the following sections.

If you do not have existing business rules, or a set of customer characteristics that you prefer to use, then we have attempted to describe a set of customer attributes; we call them characteristic variables, that all retail organizations should be capable of deriving from information captured at the point of sale. We are going to use these characteristic variables throughout this book and hopefully show you how data mining can be used to generate the business value from them. If you do not already derive these, or a similar set of characteristic variables about your customers, then you should consider doing so.

4.2 The data to be used

You clearly cannot do data mining without having the data about your customers to mine. But what data do you need? *The second stage in our data mining method* is to identify the data required to address the business issue and where we are going to get it from.

In this section we look at what types of data are typically available to retail organizations and how these can be used to determine customer characteristics. We then suggest how to construct a relatively simple data model that can be used as a starting point for discovering similar groups of customers. The data model we describe should be derivable by all retail organizations from data routinely collected at the point of sale. The model can then easily be extended to include other types of data collected about your customers.

4.2.1 The types of data that can be used for data mining

Depending on the size of the your organization, you already have a wide variety of data about your customers. This may range from transaction data, collected at the point of sale, to the storage of all types of derived information about your customers, stored in some form of data warehouse.

To perform data mining for customer characterization, there are essentially five types of data that can used to develop a customer segmentation model:

- Product data
- Transactional data
- Demographic data
- Customer relationship data
- Additional data

Although all are desirable, only the first two are essential. The five types of data are detailed below.

Note: If you already use a data warehouse, you may recognize that these data types are the typical dimensions and fact tables of a star schema, namely: Product dimension table; Sales fact table; Demographic mini-dimension table; Customer dimension tables; and if your model is a snowflake schema, additional data in Snowflake tables, respectively.

Product data

Product data is an essential requirement for data mining. The data is usually in a form that enables the Universal Product Code (UPC), usually scanned and recorded at the point of sale, to be related to the item numbers or Stock-Keeping Units (SKU) used by the retailer. This can then be mapped to the product name, the brand, pricing information, and so on. In addition, most retailers organize products into a descriptive product hierarchy.

In this type of hierarchy at the highest level, product-groups (Food, Alcohol, Household Products and so on) are divided into product subgroups (Alcohol into Beer, Wines, Spirits and so on) and each subgroup to individual product items (A-Beer(6-pack) and so on). There can of course be many more levels depending on your particular organization. Additionally, the product data may include information on when the product has been promoted or advertised and so on.

In this book we use a retail example chosen specifically to illustrate the data mining techniques. This retail organization uses a 3-level hierarchy of the type described above and the data model we describe later in this section makes use of this 3-level hierarchy. This does not preclude other types of hierarchy from being used, for example, more levels, or grouping products into departments or by brand. It also does not preclude other types of information about your products from being included in the data mining.

Transactional data

Most, if not all, retail organizations have the potential to collect and electronically store information obtained at the point of sale. If you don't already, then the remainder of this book will give you all the reasons why you should. Transactional data comprises a transaction record or sales ticket, which typically includes the following:

- The date and time of the transaction
- The store identifier
- The Universal Product Code (UPC) or other unique item code
- The quantity of the product purchased (number of items or weight)
- The price paid
- The method of payment (for example, cash, check, credit card and type, debit card)

Although it is possible to perform data mining using only transactional data, there are many advantages to be gained if the transactions can be uniquely identified to a customer. In such cases the unique customer identification number can be used to track the same customer over multiple transactions and to build up a customer profile over time. So we would suggest that **customer identifier** should be added to the list.

Ideally, the unique customer identifier should be obtained through some form of loyalty card scheme. In this case the loyalty card number not only enables the same customer to be tracked across multiple transactions but also enables other sources of data to be used.

As the name suggests, the primary aim of a loyalty card scheme is to give incentives to the customer to remain loyal usually through the reward of bonus points that can be exchanged for goods or other rewards. While this is the primary aim, the value a loyalty card brings by improving the usefulness and therefore the value of the data you collect should not be underestimated. Clearly there will always be customers who do not use a loyalty card, but by having a representative sample of customers who do, it is possible to build up a more detailed picture of your business and this can be used to benefit all your customers. In section 4.7, where we discuss how to deploy the results obtained from data mining, we describe how customers can be given incentives to use a loyalty card, or other similar means of identifying themselves during each transaction. The incentive is simply for the customer to reap the benefits that accrue from the data mining you are performing on their behalf.

There may of course be other ways in which a unique customer identification can be derived, for example, where the transaction is made by electronic payment. In this case, depending on your country's legislation, the payment card number could be used as a minimized customer identifier. However, card detail records are not routinely collected, and customers will often use different methods of payment, or different credit cards, and therefore the use of this type of approach becomes very unreliable.

Customer demographic data

Demographic data is descriptive data related to the individual customer and is not related to them being a customer. Some examples of this are:

- Age
- Sex
- Education category
- Home address
- Marital status

If the customer is a member of a loyalty card scheme, then some of this information usually can, and where possible should, be collected through the application process. While there may be a reluctance by some customers to provide this type of information, if tangible benefits of providing the information can be seen from an improved level of service, then for most people the objections will go away.

The information collected, specifically home address, can also be used as a key to unlock other sources of data obtainable from sources external to your business, for example, Consolidated Analysis Centers Incorporated data (CACI). This type of data can not only be used to enhance your knowledge of your customers, but can also be used to identify potential new customers. Data mining may also be used to discover how to acquire new customers using this type of information.

When using any type of demographic data, it is important to bear in mind the validity of the data. This will rely both on your relationship with the customer and the procedures used to capture and validate this information in your organization. In this book we do not discuss in detail the techniques for data cleansing, because this is a whole subject in its own right. Although the data mining tools that we use include statistical techniques for identifying such things as missing values and outlying in your data, these techniques cannot always remove some of the obvious inconsistencies. This type of inconsistency could be repeated records of the same customer with slight variations in the spelling of the name (for example, J. Bloggs and Miss Janet Bloggs) that a dedicated data cleansing product (for example, Vality or Trillium) will resolve. We therefore make the assumption that the demographic data used has gone through some form of data validation process.

Customer relationship data

Customer relationship data describes the data that is generated within your organization as a direct result of your customer relationship management. This type of data will typically include information on:

- Promotions already targeted at the customer
- Segments the customer has previously been assigned (either as a result of data mining or other types of customer analysis)
- Product preferences
- Channel utilization
- Complaints

Additional data

Additional data covers other data that may be useful for understanding the behavior of your customers. This data will vary depending on the retail organization but include attributes such as:

- Weather conditions
- Location of competitors and what promotions they are running
- General economic data, for example, GNP

The list is potentially endless, but if these factors are of potential relevance to the behavior of your customers, then it may be important to include them at some stage in your data mining process.

4.2.2 Suggested data models

Our objective is first to define some simple data models that can be populated from readily available data held within a typical retail organization, and second to demonstrate the types of mining function that can be performed on this data and how business value can be derived from the results obtained.

It is often incorrectly assumed that without a data warehouse containing a combination of the data types described above, it is impossible to use data mining to adequately characterize your customers. While it is true that a data warehouse will enable deeper insight of customer characteristics to be derived, significant insight can be derived from relatively simple data models.

In the following section we describe two data models that have been developed and successfully utilized in a number of retail organizations. The first data model uses only information that can be derived directly from transaction data recorded at the point of sale, and combines this with product based data using a product hierarchy. The second data model extends the first model by using a unique customer identifier to aggregate data over multiple transactions by the same customer. This second model can be easily extended to include other information on your customers.

Although these two models are relatively simple they enable tremendous insight into your customer's purchasing behavior. We show how this insight can be used to improve and develop your CRM solutions, here and in the following chapters.

4.2.3 A transaction level aggregation (TLA) data model

It is often assumed that without a unique customer identification number it is impossible to perform any valuable data mining on the transaction level data. As will be shown, while having such a unique customer number significantly enhances the value of the data mining, it is possible to perform useful analysis using individual transaction level data. This type of data mining is often referred to as basket analysis.

The minimum amount of information required to identify customer behavior is specified in the following list:

1. Transaction number
2. Date and time of transaction
3. Item purchased (identified by UPC code or equivalent)
4. Product price (per item or by unit of measurement)
5. Quantity purchased (number of items or units purchased)
6. A table matching product code to product name, subgroup code to subgroup name, and product group code to product group name
7. A product taxonomy that links product code to product subgroup code and product subgroup code to product group code.

Using this information a simple data model can be constructed. The model comprises one record per transaction with the following fields.

1. Transaction number
2. Data and time of transaction
3. Total revenue for the transaction
4. Number of articles in the transaction
5. Relative spend on each product
6. Normalized Relative Spend (NRS) on each product (see note below)

(and where a product taxonomy exists):

7. Relative spend in each product subgroup
8. NRS in each product subgroup
9. Relative spend in each product group
10. NRS in each product group

Note: *Relative spend* is defined as the proportion of total revenue spent in each category. For example, if in a transaction, a customer spends $100 on four products costing $5, $20, $25, $50, respectively, then the relative spend on each of the four products is 0.05, 0.2, 0.25 and 0.5. If the first two products are in the same subgroup and the second two products in another subgroup, the spend in each of the two subgroups will be 0.25, 0.75.

Note: *Normalized Relative Spend (NRS)* attempts to account for the significance of the purchase compared to the overall sales of the product. If in a transaction, a customer's relative spend on product A is 50%, but the sales of product A accounts for 50% of the revenue from all transactions, then customer A is exhibiting a normal buying pattern and the NRS on product A is given a value of 1.0. If product A only accounts for 10% of the total revenue from all sales to all customers, then customer A is purchasing the product five times more than the average, and the NRS of our customer for product A has a value of 5.0.

A customer level aggregation (CLA) data model

When doing customer segmentation, all data has to be aggregated to the customer level. Depending on the business practice and proposed application of the segmentation, this could be at the level of an individual customer or a household level.

Using a unique customer identification number enables the transactional data to be aggregated and combined with other data sources (for example, demographic data) and aggregated over periods of time. There are many variables that can be derived in this way and in general there will be experts within your organization who have an intuitive "feel" for which variables are likely to be good indicators of customer behavior. The following is a simplified customer model that is a useful starting point for beginning to understand your customers and which can be easily extended.

With any customer level model, the choice of time period over which to perform the aggregation depends on the stability of customer behavior. If the time period chosen is too short, then there will not be sufficient transactions to identify a normal pattern of behavior; if the time period is too long, then the customer behavior may begin to change, for example, due to seasonal trends.

The simple customer model can be derived from the simple transactional model described above by aggregating over a fixed time period to produce one record per customer with the following variables:

1. Customer ID (could be anonymous)

2. Total revenue for the customer
3. Number of transactions per customer (frequency)
4. Average time between transactions (transaction interval)
5. Variance of transaction interval
6. Customer stability index (ratio of (5)/(4))
7. Days since last visit (recency)
8. Average number of different products purchased per transaction
9. Relative spend on each product
10. NRS on each product

(and where a product taxonomy exists):

11. Relative spend in each product subgroup
12. NRS in each product subgroup
13. Fractional spend in each product group
14. NRS in each product group

Relative spend and NRS in this case are aggregates for each customer calculated for all transactions over a specified time period.

The customer stability index is a good indicator of the customers' behaviors. Low values tend to indicate customers with a regular visit pattern, while high values indicate an erratic behavior or changing behavior.

4.3 Sourcing and preprocessing the data

To create our data model we have to take the raw data that we collect and convert it into the format required by the data models. We call this stage in the process sourcing and preprocessing, and this is *the third stage in our data mining method*.

The basis for our TLA model is customer Point-Of-Sale (POS) Transaction Data combined with product data, linking product codes to product names, subgroups and so on. In the CLA model we need to tie all of this in with our loyalty card or other specific customer data. In both cases the transaction data needs to be aggregated at various levels of the product hierarchy.

If your data is stored in a relational database, then this is relatively easy to perform, because that is what relational databases are all about. If for some reason you don't use a relational database, then you have to simply resort to getting someone to code up a program to do all the hard work and create a data file to work from. You can still do the data mining on this data file, but it is just so much easier if you put all the data into the database.

Using the product code as a key into the product hierarchy and the transaction identifier as the key to each transaction, it is a relatively simple database task to perform the different aggregations required but it takes time and planning, particularly if you are talking about large numbers of customers. Typically with large volumes of transaction data involved, creating the required aggregated variables can have a big impact on your database performance and you have to think carefully about the best way to create the aggregated results you need to start data mining.

In general it is always better if you can create a dedicated table for your aggregations. Your database administrator may not like this idea and may suggest that you do it all using "views". However, since a "miner view" has to recreate the aggregations every time you run your data mining, this is potentially going to have a big impact on database performance. Another advantage of having one table is that if you really have very large amounts of transaction level data, then having the data in a large table means that you can take advantage of the capabilities to perform the data mining in parallel. This of course is only true if your data mining tool supports parallel data mining, and fortunately the one we use does.

Once you have the data model populated you are under way, but there are still a number of things that have to be done to ensure that your data mining is going to be successful. Because no data set is free from errors, there are some simple precautions that you can take to improve the mining results you produce. There are also some things that you can do to ensure you are using the most relevant information. We explain some of these steps in 4.4, "Evaluating the data" on page 63. Before we do this, we introduce the example data set that we are going to use to demonstrate the data mining techniques in this book.

4.3.1 An example data set

The data used to illustrate all of the data mining techniques described in this book is a subset of transaction records that are typical of those collected by a small supermarket. We have selected a data set comprising 500 customers who regularly purchased products over an eight week period. To aid understanding, we artificially limited the number of products from which purchases were made to

just 125 products, organized into a 3-level product hierarchy with 25 product subgroups and 5 product groups. The product hierarchy is illustrated in Figure 4-1, showing the main product groups and subgroups and some example products.

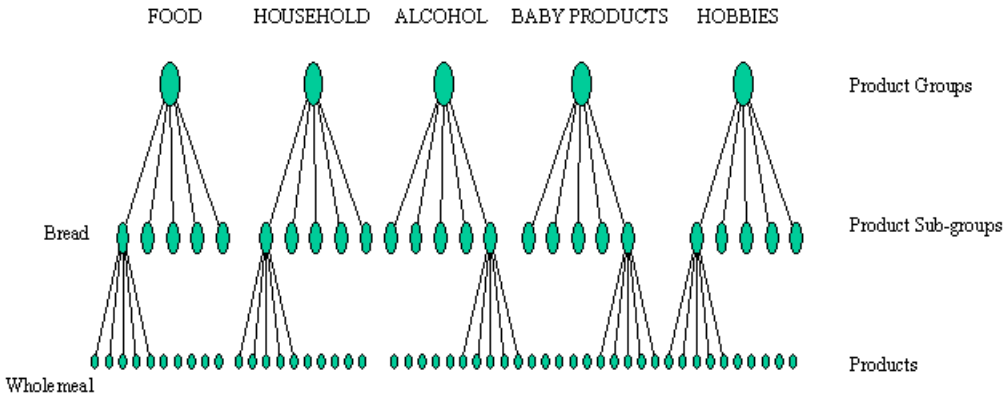


Figure 4-1 The product hierarchy

The customers were selected on the basis of how much they had previously spent in each of the five main product groups using the following predefined business rules to define the “Shopper Type”:

- General shoppers: Characterized by spending mainly on Food and Household goods with a relatively small expenditure on Alcohol and Baby Products and Hobbies. Fifty percent of customers are in this group.
- Family shoppers: Characterized by spending mainly in the Food, Household and Baby Products groups with a small expenditure on Alcohol and Hobbies. Twenty-five percent of customers are in this group.
- Affluent shoppers: Characterized spending in four of the five categories, particularly Alcohol and Food but with a small expenditure in Baby Products. Ten percent of customers are in this group.
- Alcohol and Hobby shoppers: Characterized by having most expenditures in these two categories. Ten percent of customers are in this group.
- Hobby shoppers: Characterized by purchasing predominantly from the Hobbies category, with little or no expenditure on all other products. Five percent of customers are in this group.

The initial data set comprised 5000 transaction records representing an eight week period in which in excess of 50,000 individual items were purchased. This data was used to populate both the TLA and CLA data models. The data mining tool that we use enables us to graphically display the distribution of the variables from our data models, for each of the shopper types. Figure 4-2 shows the bivariate statistics for the CLA data model by grouping on the “Shopper type” variable and displaying the NRS aggregated at the product group level.

Note: Throughout this book we are going to be looking at a lot of diagrams similar to Figure 4-2 through Figure 4-4, and so it is important that you understand how to interpret them.

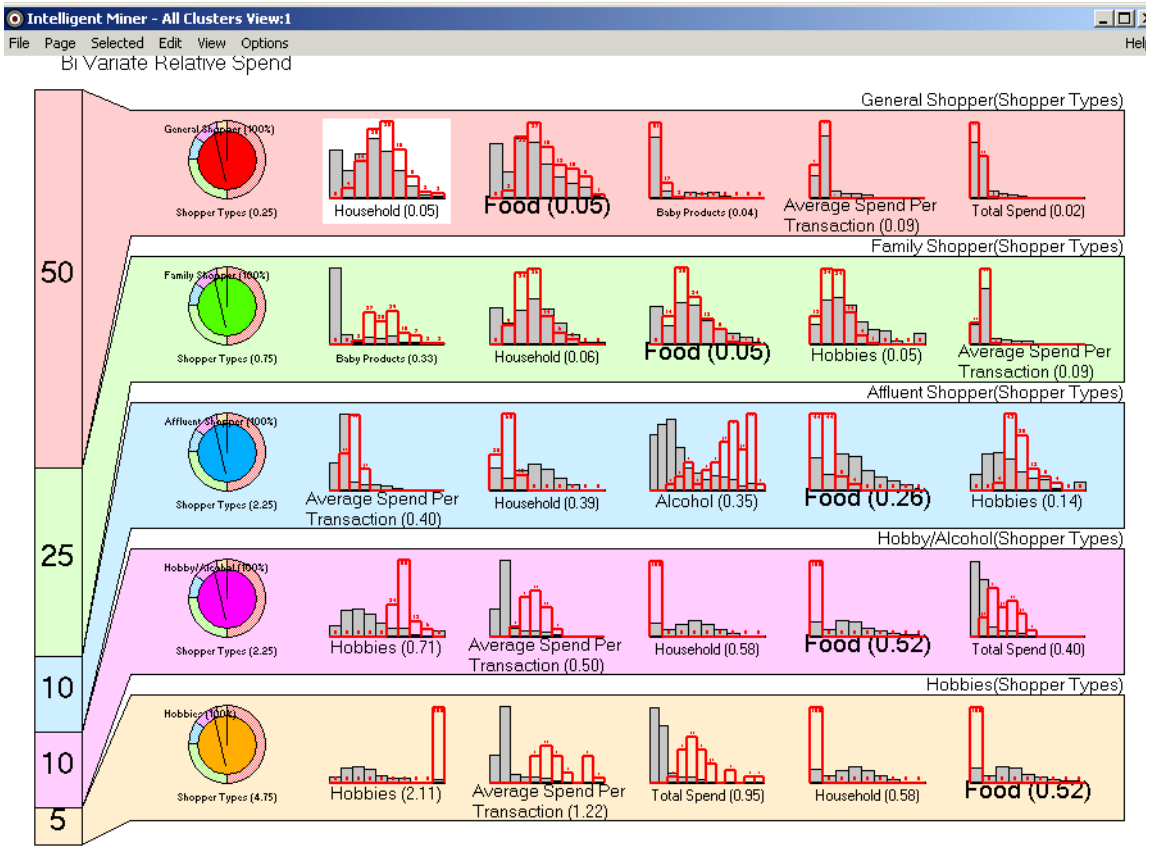


Figure 4-2 Distribution of variables for the different shopper types

In Figure 4-2 the data for each of the five Shopper Types is shown on a separate line with the name of the Shopper Type indicated on the top right hand side of each line. If we think of each of our Shopper Types as market segments then in this type of display, the percentage of customers in each segment is shown on the left hand side. The distribution of each of the variables from the data model is depicted either as a pie chart (for categorical variables) or as a histogram (for continuous or numeric variables). Only six of the customer variables from the CLA model are visible at this level of the display. If we want to see all of the variables for a particular segment, we can select one of them (for example, the Family Shopper segment, 25%) and drill into the segment. This is illustrated in Figure 4-3 where we show eleven of the variables from the CLA model.

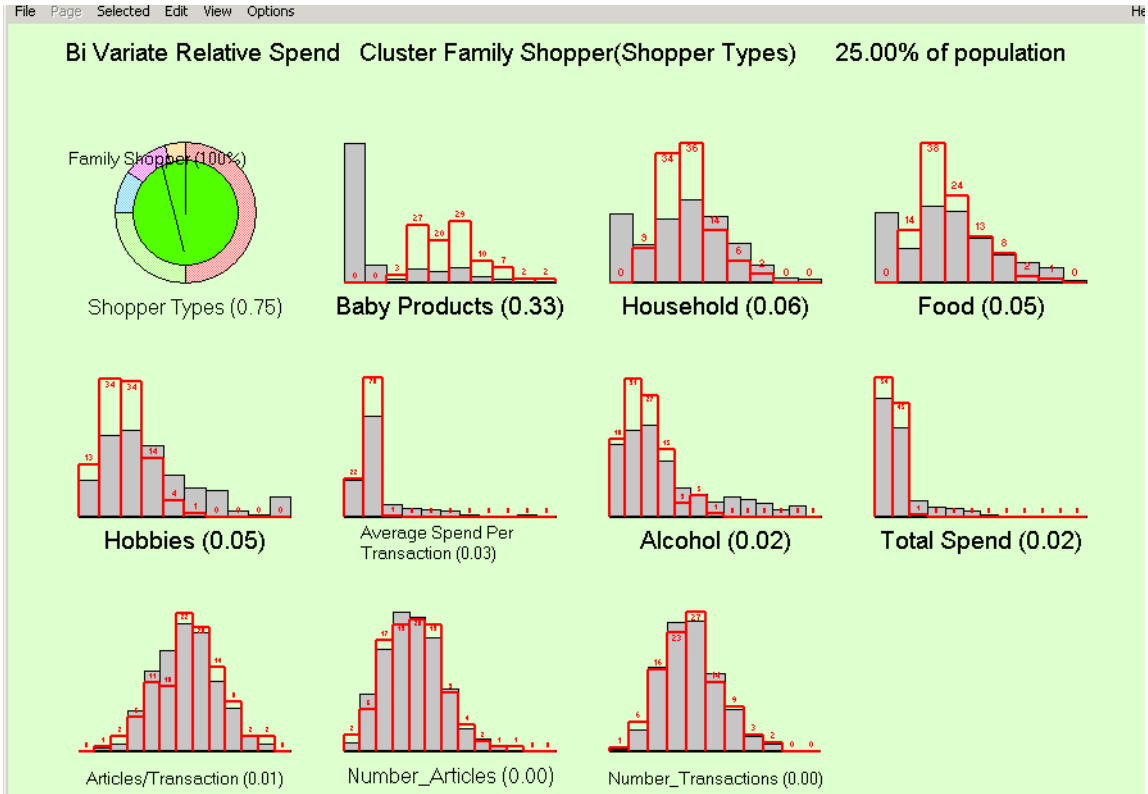


Figure 4-3 Distribution of variables for the General Shopper segment

In our data model we have only one type of categorical variable, the “Shopper Types”. The pie charts shows the distribution of this variable within the segment (inner circle). In this case, because we have selected the segments using the Shopper Type, the pie chart confirms that we have only one type of shopper in this segment and is labeled “Family Shopper (100%)”. The outer annulus of the

pie chart shows the distribution of the different shopper types in our total data set, and this tells us that 50% of the records are of one type (General shopper) and that 25% are of the type Family Shopper and so on. This enables us to understand how the shopper types in this segment compares to the distribution of the variable for all customers. The number in brackets against each variable is a statistical measure of how different the variable for the particular Shopper Type is in comparison to the distribution of the variable for all Shopper Types. This is explained further at the end of this section.

The histograms show the distribution of the continuous variables from the CLA model for customers in this segment. If we concentrate on the Baby Products product group, then the non shaded histogram shows how the Family Shopper customers normalized spend on baby products is distributed. This can also can be compared with the distribution all our customers spend on baby products, which is the shaded histogram.

Again we can drill into this diagram to get a clearer picture of this distribution as shown in Figure 4-4.

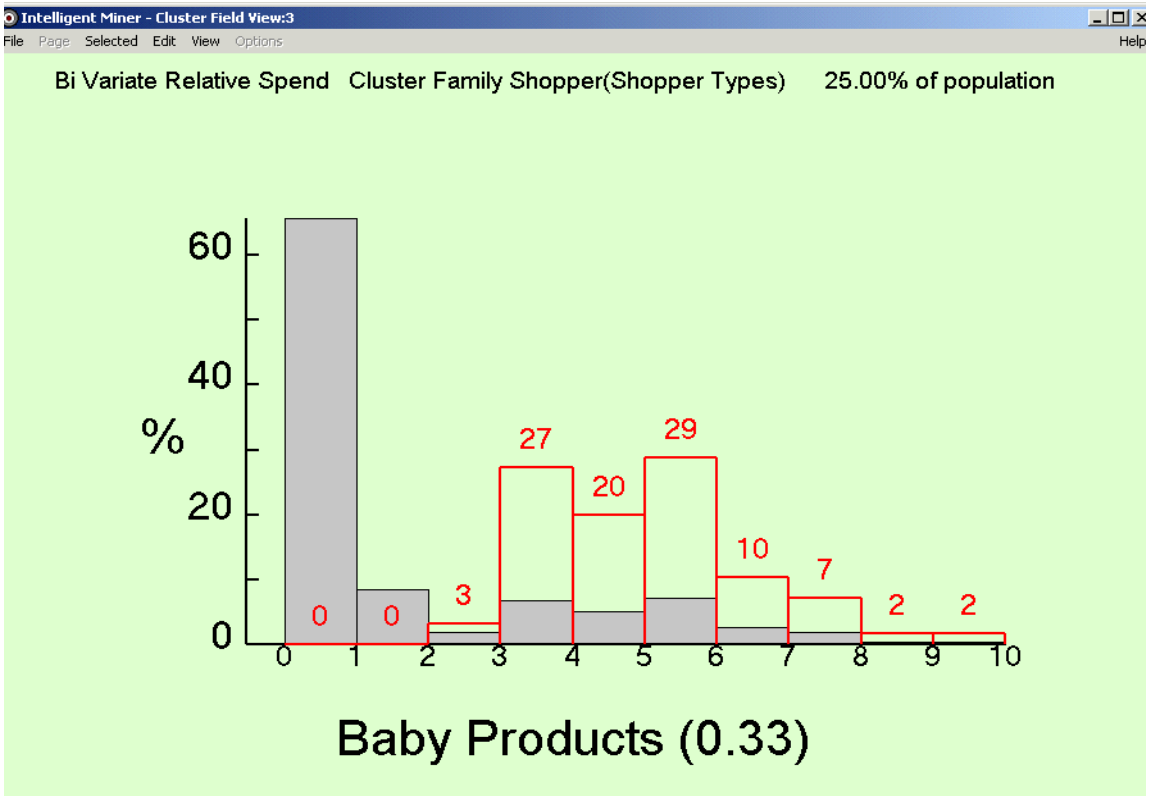


Figure 4-4 Family Shoppers normalized spend on Baby Products

This shows us that Family Shoppers normalized spend on baby products is between 2 and 10, and the normalized spend over this range is much higher than that for all customers. For example, 29% of customers in this segment have a normalized spend of between 5 and 6 (the value being shown above the column), which compares with 7% for all customers. This should not surprise us, because normalized spend itself is a measure of how much more a customer spends on a particular product compared to all customers.

Returning to Figure 4-2, you will also see that the customer variables in each segment are arranged in a different order. The order is important since it tells us which are the most important variables to concentrate on when we want to describe the segment. To do this the data mining tool has associated, with each customer variable, a number which it uses as a sort value. This number is called the Chi-square value and can be seen in Figure 4-3 next to the variable name (see note below).

Note: Without going into detail the *Chi-square value* is a statistical measure of how different the distribution of the variable is for customers within the segment, compared to all customers. So for our family shoppers the spend on Baby Products is a good indicator that they are in this group, followed by their spend on Household goods, then Hobbies and so on. If the Chi-squared value is zero for a variable, then the distribution of the variable for customers within the segment is no different from the distribution for all customers. We can see this in Figure 4-3 for the variables “Number of Articles Purchased” or “Number of Transactions”, where the value is almost zero and our two sets of histograms are almost identical.

4.4 Evaluating the data

Having created and populated our data models, *the fourth stage in our data mining method* is to perform an initial evaluation of the data itself. In 3.4, “The generic data mining method” on page 29, we looked at the sort of things you have to think about before you begin to mine your data, and in this section we apply this to our example data. We usually do this in three steps.

4.4.1 Step 1 — Visual inspection

In the previous section we started to look at our example data using the visualization tools available to us. In this case we had the predefined business rules to guide our selection, and so we were able to group the data according to the Shopper Type and make some initial interpretation. However, suppose we did not have any predefined business segments, then the first step would be to look at the data for all of our customers. Figure 4-5 shows us what we might typically be confronted with.

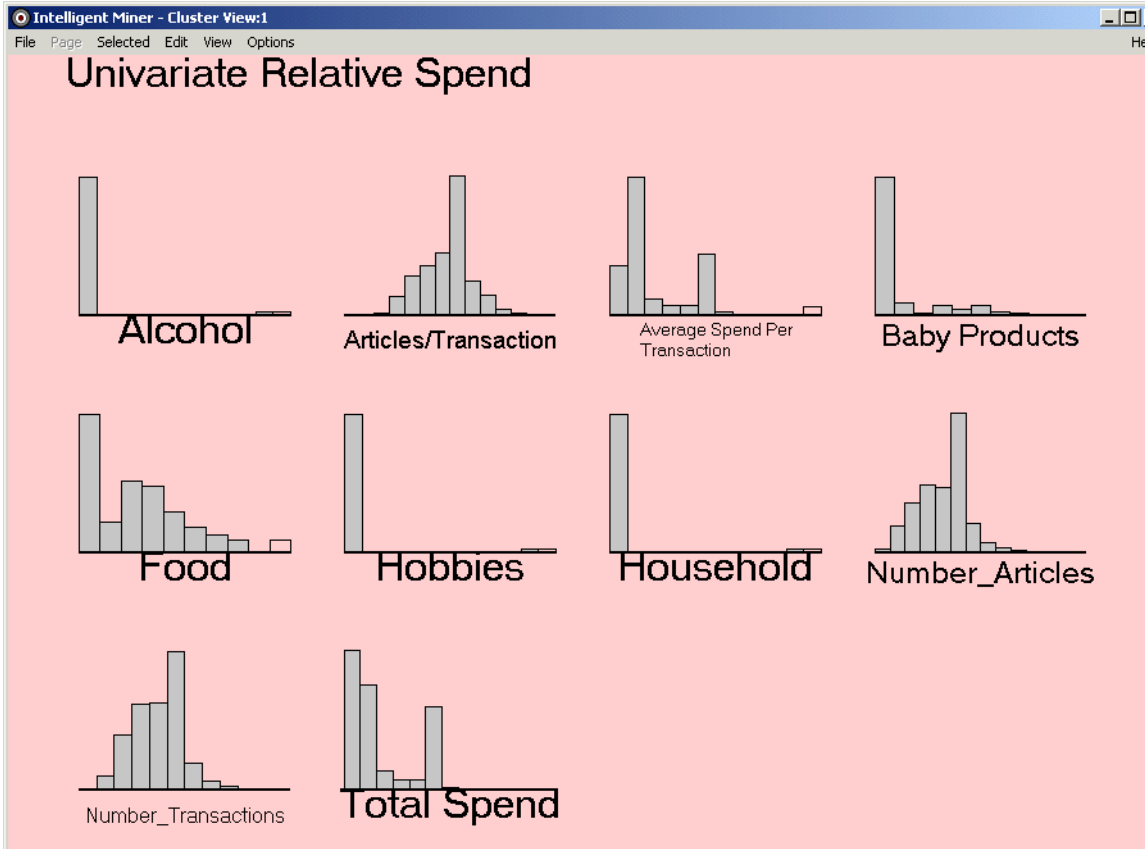


Figure 4-5 Distribution of variable from the CLA model for all customers including outliers and missing values

The picture looks somewhat different because we now have the aggregated data in each variable for all of our customers. We know that buried within this data are customer segments similar to those we used when selecting the data. The question is can we rediscover the segments using our data mining tools and techniques?

However, before attempting to do any data mining it often pays dividends to just visually inspect the data and ask ourselves the question “Does it look right?”. The reason we do this is quite simple. While there are a number of statistical tests that we can perform on the data and use these tests to determine all manner of things about the distributions and correlations that exist, statistical tests cannot apply business experience. There really is no substitute for using your business knowledge and experience to make an initial assessment of the validity of the data you have. As an example, it is very unusual to find a set of data that is

completely free from errors and often a simple visual inspection will tell you that things are not what you expected. Obvious errors can be spotted very quickly and either removed from the data or accounted for in your analysis. In the case of our example data, we can immediately see in Figure 4-5 that we have some unusual distributions for our customer variables, particularly for the Alcohol, Hobbies, Household categories. On closer inspection we would discover that there are some outlying values that are biasing the histograms. These could be a real effect, but you don't need a data mining tool to reveal this to you, and if they are real then you may have already discovered something about your data that you did not know. In this case, the outliers are due to data input errors and we can remove them.

There are all sorts of statistical techniques that can be used to identify less obvious inconsistencies in your data, and without going into detail we describe the sorts of things you have to consider.

4.4.2 Step 2 — Identifying missing values

One of the most important things to identify are missing values for some of the variables you are proposing to use. Missing values can introduce bias into the results and you need to be made aware of their existence. The data mining tool that we use automatically identifies and reports missing values (for example, the non shaded histogram column on the right hand side in the Food variable in Figure 4-5 shows the missing values in this variable).

The data mining techniques can take some account of missing values, but there are a number of ways in which this can be done and we have to decide the most appropriate action. In the case of our example data set, we decided to remove both outliers and customer records with missing values.

Just doing this type of analysis and taking some simple steps to remove obvious inconsistencies can give a much clearer picture of our data as shown in Figure 4-6.

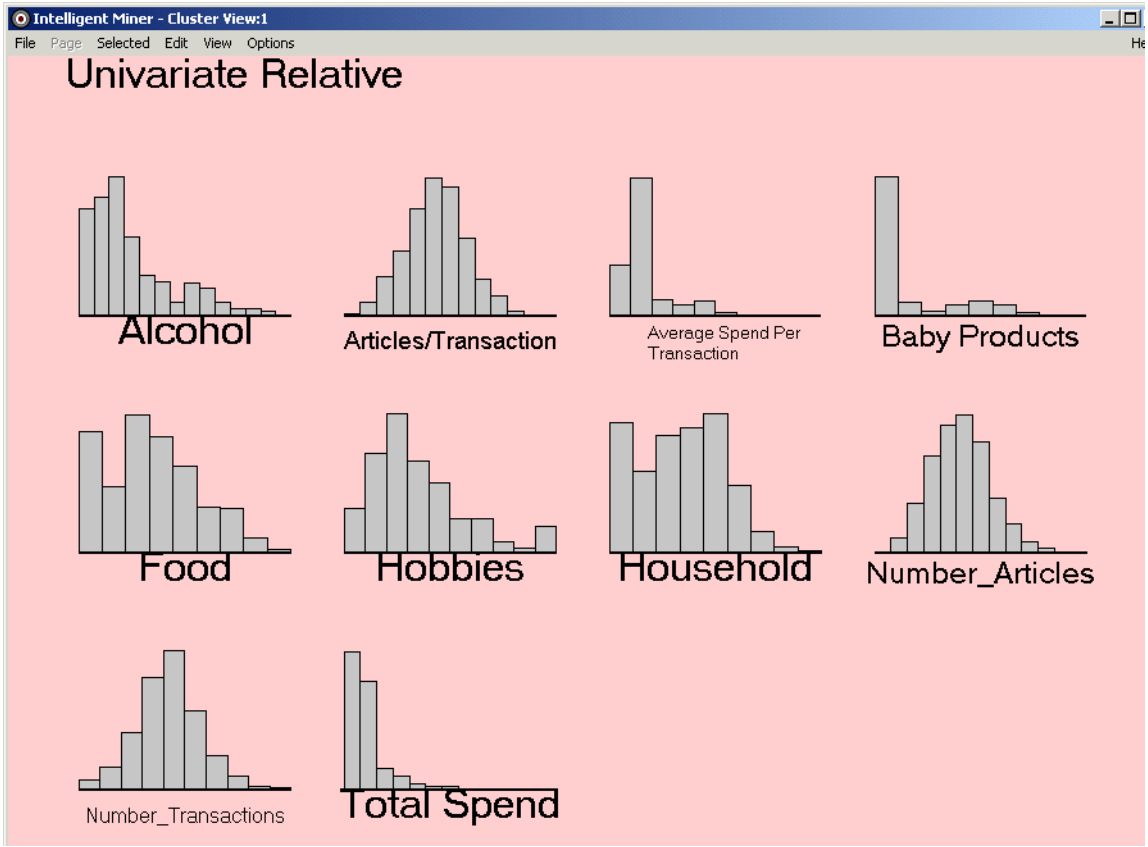


Figure 4-6 Distribution of variable from the CLA model for all customers after having removed the outliers and not showing the missing values

4.4.3 Step 3 — Selecting the best variables

To illustrate how to use the data mining techniques we have used the CLA model and have chosen to aggregate the transaction data at the product group level for each customer. In the case of our example data, there are only 10 variables. In most retail organizations there will be many more product groups (typically ~100), and therefore even with the CLA model there will be many more variables to consider. One of the first things a data miner tries to do is to reduce the number of variables to a meaningful subset. The reasons for doing this are:

- To produce good results
- To produce results that are easily interpreted

Variables that give good results

To produce a good results from data mining we need good customer characteristic variables. This does not mean that we want to pre-judge the data mining and only select the variables that we think are going to be useful; the data mining can do that for us. What we mean is that although we may not know which variables are important, one thing we can say with certainty is that if two variables are strongly related to each other (for example, the day of the week expressed as a number and the name of day) by including both of them in our data we increase the number of variables, but we did not add any extra information that we can use. We talk about such variables as being highly correlated. Removing highly correlated variables is therefore very desirable. If we know that some variables are just expressing the same information in another way, then we should choose the one that is most descriptive to us. Using standard data models like the TLA and CLA will of course mean that someone else has thought about this for us, but if we extend these models (with demographic data, for example) we need to consider how these variables might be related.

Even using standard models does not solve everything. The reason for this is that when we populate the model with data, the variables can still be strongly correlated. This is simply because things are not independent. If our customers always purchased food and baby products in equal proportions, such that if we doubled our expenditure on food, we doubled our expenditure on baby products, then we have highly correlated variables. In other words, we can predict what will happen to one variable as we change the other.

Note: We can perform a number of statistical tests to find out how correlated our variables are (for example, bi-variate statistical analysis, linear and polynomial regression). We can even create new variables that combine the correlated variables into a one single variable (for example, principal component analysis and factor analysis). Whatever techniques we use, the redundant variables should be identified and we should then choose which ones we want to use.

This all sounds very difficult, and in many cases you can do useful data mining without doing all this advanced analysis. However, if we take the time to evaluate our data in this way, then it will pay dividends and we will get a much better result.

Variables that produce interpretable results

Data mining is about discovery, about finding things about your business that you did not previously know. That's one of the things that makes it different from simply using statistics to confirm or rebuff what you already knew or thought you knew about your business. Producing a correct result is of course important,

indeed essential, but if we can't interpret the result we get, then we do not know what we have uncovered. To produce an easily interpretable result, we therefore need to ensure that we choose variables not only in the sense that they are statistically correct, but that they make sense to the people who have to use the results.

Using a standard data model again helps, but if there a lot of variables, then even when we get a result we might find it difficult to understand. This is a bit like our discussion of business rules. We can have lots of customer segments defined by different business rules, but if we have too many variables our ability to interpret the results becomes the challenge. Data mining can deal with many more variables than you can keep in your head and can present the results to you in as many of these variables as it takes, but if you can't understand the result, it is of questionable business value.

Note: There is an important distinction to be made between segments produced by data mining and those produced by business rules. As the business rules become more complex it gets harder to ensure that customers who match one rule do not also match the others. In other words, they are no longer exclusive. Data mining segmentation preserves the exclusiveness, and therefore we are able to write detailed rules that uniquely define a group of customers. This is one of the reasons data mining can find niche markets where business rules cannot.

You also have to make use of the results in your business, and discovering new things about your customers without the ability to react to what you have discovered is of little value. Sometimes it is better to accept a result that you can understand and act upon, rather than striving for the ultimate performance but lose the ability to interpret the result. Practical experience of data mining in many retail organizations has shown that sometimes the "gut feeling" of business users is often as good an indicator of what variables to use, than the most sophisticated statistical analysis.

Hint: One simple way of reducing the number of variables and at the same time making the results easier to interpret, can be used where a predefined business customer segmentation exists. The technique makes use of the Chi-square statistic that we describe above. In this case the variables can be sorted as we showed in Figure 4-2, and then we can select those variables with the highest Chi-square values from all segments. Doing this will produce a data driven segmentation that is closest to the business rules segmentation, provided the data supports the business rules. This is a good starting point for data mining, since the results will either reinforce your business view by producing the same segments, slightly modify your view by suggesting similar but usually more segments, or drastically modify your understanding of your customers by producing completely different segments (our experience is that the second of these is the most usual.

Having said all that, you still might not know which variables to choose, for example, by using our data models, there could still be a large number of product groups and they may all be important. Fortunately as we will see, the data mining techniques can help you to make this selection.

4.5 The mining technique

Choosing the mining techniques to use is *the fifth stage in our generic mining method*. The mining technique we use to identify customer behavior is the segmentation or clustering technique. Clustering is a discovery data mining technique. What people usually mean when they talk about discovery data mining is that it does not require any prior knowledge of your customer segments to make its own decision. There are a number of different types of clustering techniques that can be used. In this next section, we consider two very different types and explain how you can decide which is the most appropriate one to use and how it should be applied.

4.5.1 Choosing the clustering technique

The data mining technique of clustering is a process that attempts to group together similar customers based on the variables that you have measured, while at the same time seeking to maximize the difference between the different types of customer groups it forms. Different, but very homogenous clusters tell a specific story about customer behavior, for example, preferred products or product combinations. Provided care is taken in selecting the customer variables (as we described above), all clustering techniques will, if used correctly, produce groupings of your customers. Unfortunately, other than in simplest cases, it is rare that two clustering algorithms will produce exactly the same groups. It is

important to understand that **different clustering techniques produce different views of your customers**. This may come as a shock — after all isn't data mining supposed to find the clusters for you? But it should not be such a surprise.

Suppose you asked different people to segment your customers. You would expect to get a number of different views and opinions. Indeed the more customer variables you provided with which to make the decision, the more opinions you would get. Its just the same with clustering. Different clustering techniques take a different view of what is the best. However, just like people, some opinions are more valid than others and some clustering techniques produce better clusters than others. The challenge is to determine which view is the best. The main reason for the different views, at least in the case of clustering techniques, is the definition of what we mean when we say that we are trying to group similar customers together, while at the same time keeping them separate from customers who are different. What exactly do we mean by similar and different and how do we measure this?

If you asked a number of people to group your customers, most of them, but not all, would ask, "How many groups do you want?". Most clustering techniques do the same, they require you to specify how many clusters you want and then they will group customers into the number you specify. Even so, neither the people you ask nor the clustering techniques will end up with exactly the same view. Of course, there may be some customers groups that are so obvious that however you group your customers these groups will always be chosen, but this is rarely the case, particularly for retail data. So what is the solution? One possible solution is to get a consensus. Get the views of more than one and find out where they agree and disagree. This is precisely the approach we use.

The two clustering techniques that we describe are called demographic clustering and neural clustering.

Demographic and neural clustering

Demographic clustering, as the name implies, is a clustering technique that was developed specifically to cluster demographic data. Since demographic data contains large numbers of categorical variables, this clustering algorithm works very well with this type of data. When it has to cluster using continuous variables, it does so by treating them almost as if they were categorical.

The Neural clustering technique does almost the reverse. It works well with continuous data but treats categorical data as if it were numeric (0 or 1) for every categorical value. Trying to claim that one is better than the other depending on the type of data is just a generalization. Because they are approaching the problem from different angles, both techniques can be used to complement each other and to gain confidence that the data driven segmentation is indeed producing a valid result.

In the following sections we will describe how to obtain clusters using the two clustering techniques and then consider how to interpret and rationalize the different views that they give of your customers.

4.5.2 Applying the mining technique

To the user of these two clustering techniques, the most obvious difference is that the neural clustering technique needs you to specify how many clusters you want, while the demographic clustering does not.

Demographic clustering

Demographic clustering will try to discover the number of clusters automatically. To do this however it requires you to specify how similar the customers in any one cluster should be before they are they can be grouped together. This is simply a question of you specifying a threshold value between 0.0 and 1.0 where a value of 1.0 means they are identical and 0.0 means that they can be completely different. If two customers are more similar than the threshold value, then they are candidates for being put into the same cluster (see note below).

Note: *Similarity* between two customers is calculated by comparing each customer attribute variable and giving a score for how closely the variables match. The scores are then summed and divided by the number of variables you compared. If all the variables are categorical then this is easy to understand. Take gender, for example. If you compare two customers who are both male, then gender attribute variables are identical and they contribute a score of 1.0; if they are different (male and female), they contribute a score of 0.0. So if a customer is described by 10 categorical variables and we compare two customers, if any five variables are the same and the other five are different, then the two customers are 50% similar and could be grouped together.

For continuous variables the concept of being the same is slightly different. If the values are identical, then they contribute a score of 1.0, but if the values differ then they get a score based on the degree of difference. Typically, if the difference is expressed in terms of the number of standard deviation of the variable for all customers and the score is calculated such that if the two values are 0.5 of a standard deviation apart, the score is 0.5. Although standard deviation is normally used, other measures can be defined. If the similarity threshold is 0.5, then customers can potentially be grouped together when the sum of the scores for each variable divided by the number of variables is greater than 50%.

Clearly if we specify a similarity threshold of 1.0, then we are insisting that to group customers their characteristics must be identical. If we have a large number of customer characteristic variables or the variables can take a wide range of values then we would probably end up with each customer being assigned to there own unique segment and we would have achieved perfect CRM. At the other extreme if we set the similarity threshold to be zero then we end up with one large cluster containing all of our customers since all of them become candidates. The answer is somewhere in between but where? The balance has to be between the number of clusters that is acceptable and the degree of similarity.

The other important factor that the clustering algorithm has to consider is that even if a customer has an acceptable similarity to an existing group of customers, this does not automatically mean that they will be put into the same cluster. There may be other customer groups where the match is better. The demographic clustering technique finds optimum combinations of customers that maximize their similarity within each cluster, while at the same time maximizing the dissimilarity between different clusters. To decide that this it tries to maximize the value of a statistic that it calculates called the Condorcet value.

Note: *Condorcet* like similarity has value in the ranges zero to one and is a measure of how similar a customer is to other customers within the cluster, and how dissimilar they are to customers in other clusters. It has a value of 1 when all customers in a cluster are identical and where there are no customers outside the cluster that have the same characteristic. A value of zero indicates that the customers are distributed randomly among the clusters. The condorcet value can be calculated for all the customer variables or for each variable separately.

If we give the demographic clustering algorithm a similarity threshold and do not limit the number of clusters that it can produce, it will keep trying to find the minimum number of clusters that satisfy the similarity threshold since this will also maximize the condorcet value. If we do not know what similarity threshold to select, there are some tricks that can be used to discover what the optimum value may be, and we describe this below.

Without going into details, using our statistical analysis techniques, we were first able to determine that the following variables could be removed from the analysis:

- Normalized relative spend on Food
- Normalized relative spend on Household
- Total Number Articles Purchased
- Total Number Transactions
- Number of Articles per Transaction

Having removed these variables and clustering with an initial similarity threshold of 0.5, gave us the clustering result shown in Figure 4-7.

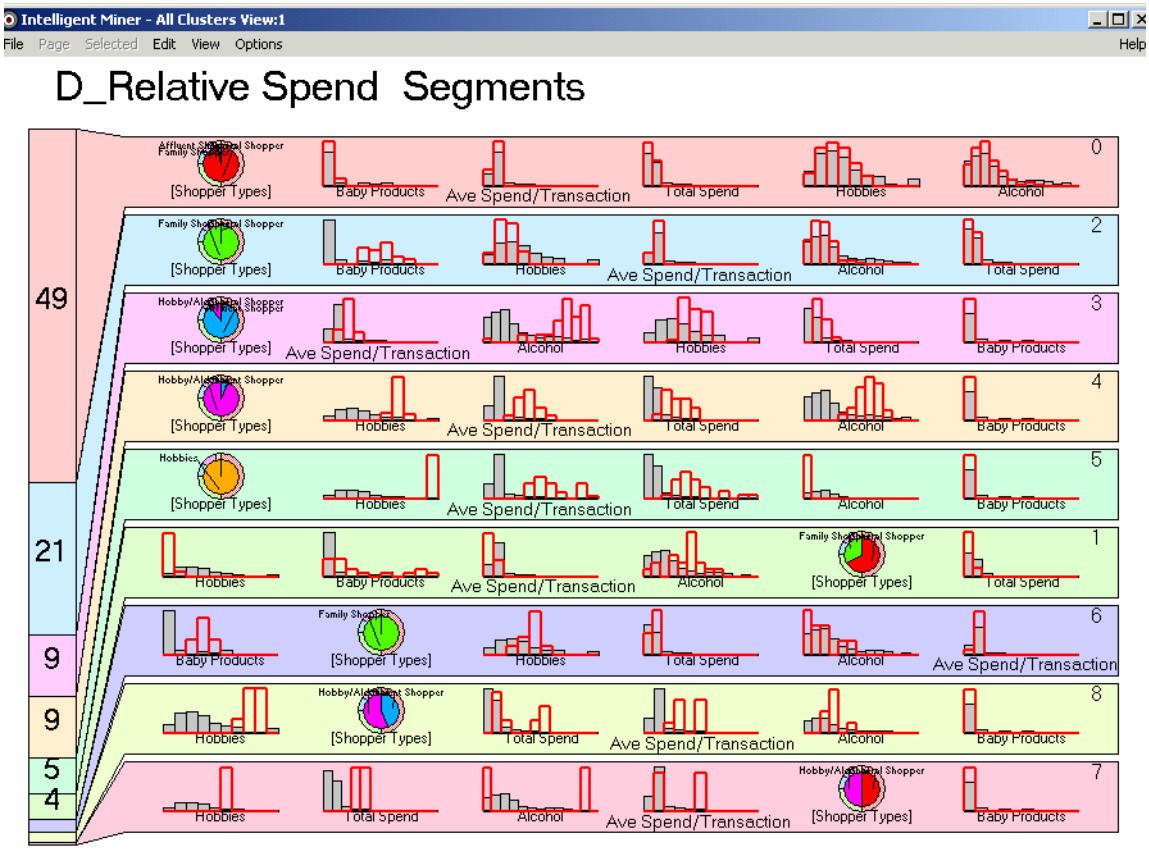


Figure 4-7 Demographic clusters for a similarity threshold of 0.5

The cluster result is presented in a similar way to the results we looked at for the business rule customer segments. In this case however, each line is now a cluster determined by the algorithm, ordered from top to bottom by the cluster size (again indicated by the number on the left hand side). The cluster is identified by a unique cluster number shown at the top right hand corner of each line. Again the variables are ordered from left to right according to their statistical significance and we have included some supplementary variables (which we must emphasize were not used to generate the clusters) in the results.

The demographic clustering technique suggests that there are 14 clusters ranging in size from 49% of our customers to some very small clusters (<3% of customers in each). Although we have not used the Shopper Type to create the clusters, there is a good indication that the automatically generated clusters are an excellent match to our initial business rule segmentation. The largest cluster,

for example, is comprised almost entirely of the General Shopper Type and the second largest almost entirely of the Family Shopper Type, and similarly for the next three segments for the other shopper types. The size of the clusters are also what we would expect. The smaller clusters are then a mixture of shopper types.

This is a very encouraging result and seems to confirm that we can indeed extract correct clusters from our initial data. However, judging the quality of the cluster does not depend on how well it matches the business rule segmentation (after all, we may have gotten it wrong), but on how good the clusters themselves are in terms of the Condorcet value for the clusters and the number of clusters produced. In this case the value is 0.62 which is marginal (a minimum value ~ 0.7 is usually acceptable), but not totally unexpected for retail data. However, at this value of similarity, experience has shown that there will often just be one large cluster and a number of small clusters; so without the benefit of having the business rule segmentation to compare with, we would not know whether this was an acceptable result or not.

Again, without going into too much detail, it is possible to investigate how the Condorcet value changes in response to increasing the similarity value. To do this and since we did not want to have a large number of small clusters, we fix the maximum number of clusters at nine and found that the Condorcet value was maximized by choosing a similarity threshold of 0.8. The result is shown in Figure 4-8.

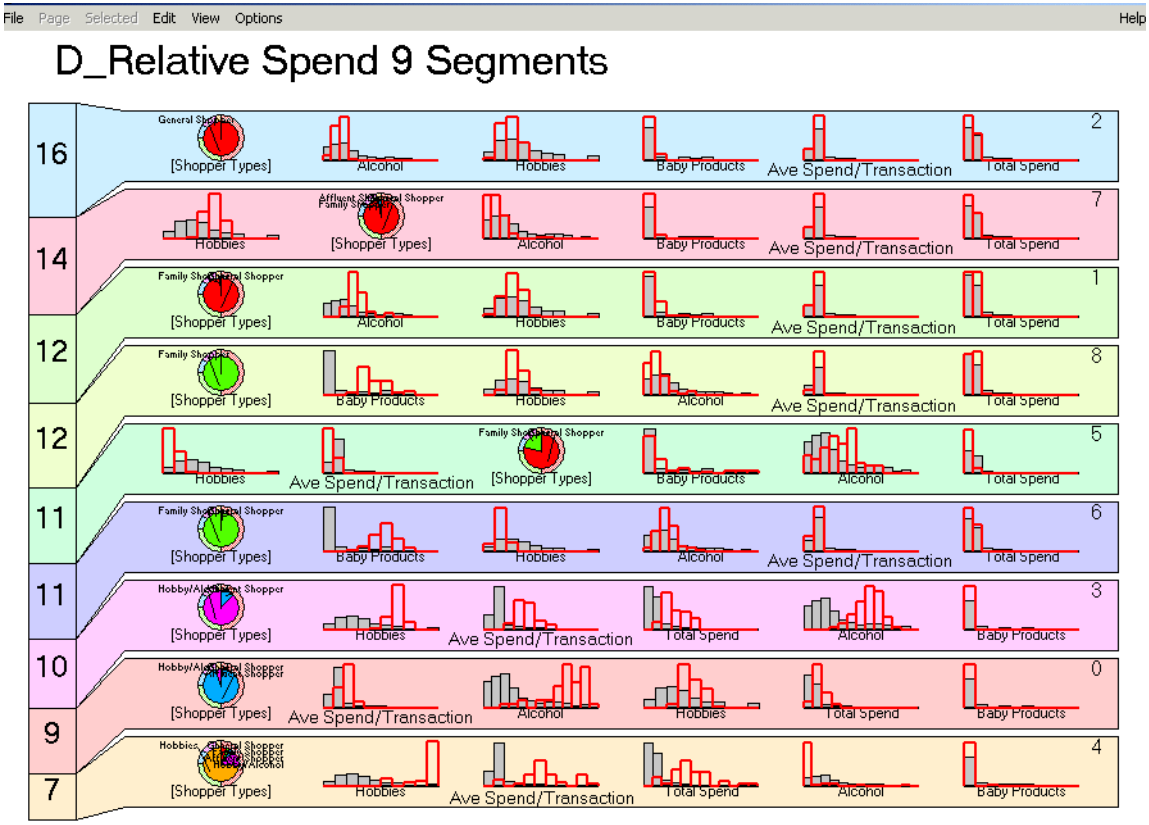


Figure 4-8 Demographic cluster at a similarity threshold of 0.8

We now have more even sized clusters with the an overall Condorcet value of 0.69. In this case, we now have the Shopper Types dispersed across a number of clusters:

- Three clusters of General shoppers (Cluster 2, 7, and 1)
- Two clusters of Family shoppers (Cluster 8 and 6)
- One mixed cluster of General and Family shoppers (Cluster 5)
- One cluster of mainly Hobby/Alcohol shopper mixed with some Affluent shoppers (Cluster 8)
- One cluster of predominantly Affluent shoppers (Cluster (0)
- One cluster of mainly Hobby shoppers (Cluster 4)

Initially, it may seem to be a poorer result than the one obtained at a similarity threshold of 0.5. However, as we will see when we come to discuss how to interpret the results, this is a more interesting segmentation of our customers. It is also a segmentation that we can justify from a statistical analysis, rather than on a subjective comparison, with our business rule segment, which may be incorrect.

Neural clustering

Having determined by using the demographic clustering technique that using nine clusters produces a valid segmentation of our customers, it is relatively easy to use the neural clustering technique to generate the same number of clusters, using the same variables.

As we explained earlier, the neural clustering technique does not require us to specify things like similarity, but only the number of clusters we want. In addition however, the neural clustering technique also requires a parameter that tells it how long we are prepared to wait before it gives us the result (called number of passes). We forgot to mention that people who ask you how many segments you want, also ask you how long do they have before they have to give you an answer. If you are just using the neural technique, then you have to experiment with the number of passes to determine the stage at which it produces a set of clusters that are stable (not changing as you increase the number). In our case the number was 16 passes. The results are shown in Figure 4-9.

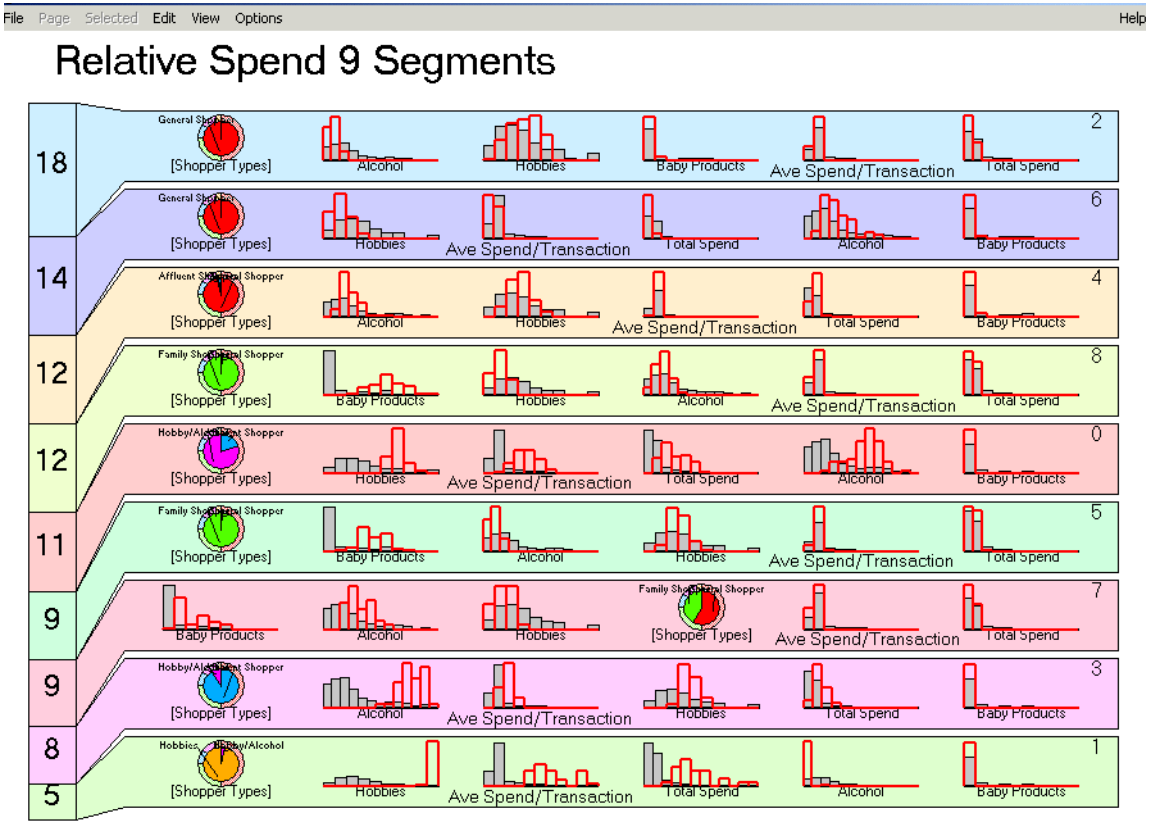


Figure 4-9 Neural cluster result for 9 clusters (3x3)

The results are presented in exactly the same way as the demographic cluster results, although as we will see in the next section there is a significance to the unique cluster number that is assigned. Summarizing these results using the business rule Shopper Type we have:

- Three clusters of General shoppers (Cluster 2, 6, and 4)
- Two clusters of Family shoppers (Cluster 8 and 5)
- One mixed cluster of General and Family shoppers (Cluster 7)
- One cluster of mainly Hobby/Alcohol shopper mixed with some Affluent shoppers (Cluster 9)
- One cluster of predominantly Affluent shoppers (Cluster (3)
- One cluster of predominantly Hobby shoppers (Cluster 1)

A simple visual inspection of Figure 4-8 and Figure 4-9 using the Shopper Type as a guide seems to show that other than for the cluster number assigned and the order in which they are arranged in the two output visualizers, the two sets cluster results are very similar.

However, on closer inspection we will see that this is somewhat superficial and that the segments produced are not as similar as this simple analysis would suggest. So what is happening and how do we understand what the two sets of results really mean? In the next section we explain it all.

4.6 Interpreting the results

In the previous section we looked at the steps we have to follow to get our mining results using two different clustering techniques. The *sixth stage in our generic mining method* is to interpret the results that we have obtained and determine how we can map them onto our business. When you are first confronted with the cluster results the first question that you are going to ask is “What does it all mean?”. In this section we describe how to understand and read and interpret the results from the different clustering techniques, but more importantly how you can compare the results from different cluster techniques.

4.6.1 How to read and interpret the cluster results?

The cluster techniques that we have used both produce results that can be displayed graphically, as we have seen. We can also obtain additional visual information by highlighting individual clusters and even individual variables. There is also another level of detailed information which gives us the statistical information that we need to fully interpret what the clusters are telling us about our customers. Although the visualized results are important in giving us an overall impression of what is happening, the interpretation always needs to be backed up with the statistical detail to confirm our understanding. In this section we look at how this is done.

As we discussed in 4.2, “The data to be used” on page 49, the first thing you need to understand about the visualized results is that the graphs and charts are telling you about the characteristics of customers who have been put into the cluster and how these customers differ from the population as a whole. It is not telling you directly how the customers in one cluster differ from customers in another cluster. This is an important distinction and we will return to this issue later.

When we describe our clusters therefore, we are looking at the characteristic variables of our customers in each cluster, and using these to describe how customers in this cluster differ from the customers in all other clusters. To do this we want to choose those variables that make the greatest difference, and fortunately the order in which the variables are presented to us (using the chi-square statistic) give us just what we need.

Figure 4-10 shows the expanded visualized result for Cluster 8 from our demographic cluster results. In expanding the view we have also included some of the characteristic variables that were not used to produce the cluster, but these help us to interpret the results.

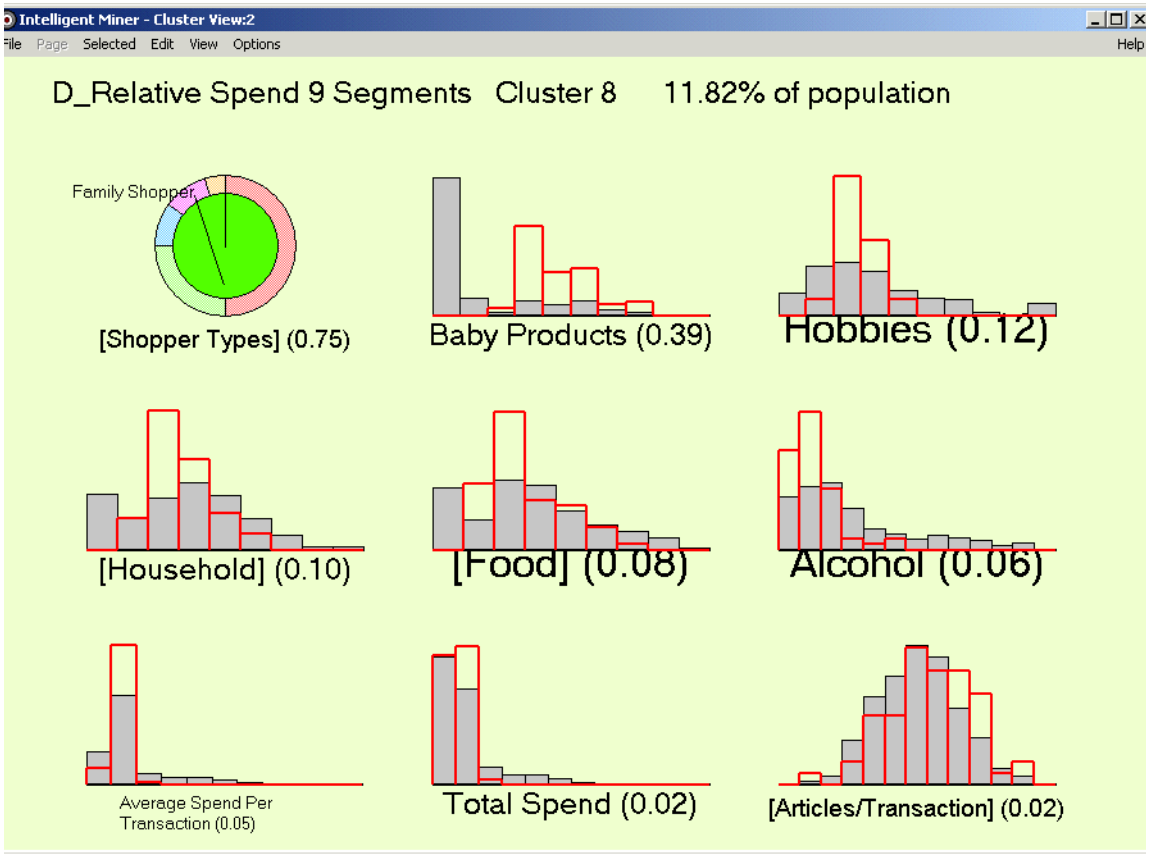


Figure 4-10 Demographic cluster 8

This cluster is the first of the two clusters that contained what our business rules defined as Family Shoppers. You will remember that the business rule for this group was:

- Family shoppers: Characterized by spending mainly in the Food, Household and Baby Products groups with a small expenditure on alcohol and hobbies. Twenty-five percent of customers are in this group.

The first thing to note is that this segment represents 11.82% of our customers and so we have about half of the customers we previously characterized as Family Shoppers in this group. The next thing that we note is this cluster is determined predominantly by the spend on baby products and then on hobby products, but now the sort value is becoming much smaller as there is less to distinguish the spending of this group of customers from customers in the other clusters.

If we look at the detailed view of the corresponding shopper type cluster (Cluster 6) shown in Figure 4-11, we can again see that the cluster is determined by the spend on baby products, but now there is a greater spend on baby products and less is spent on hobbies.

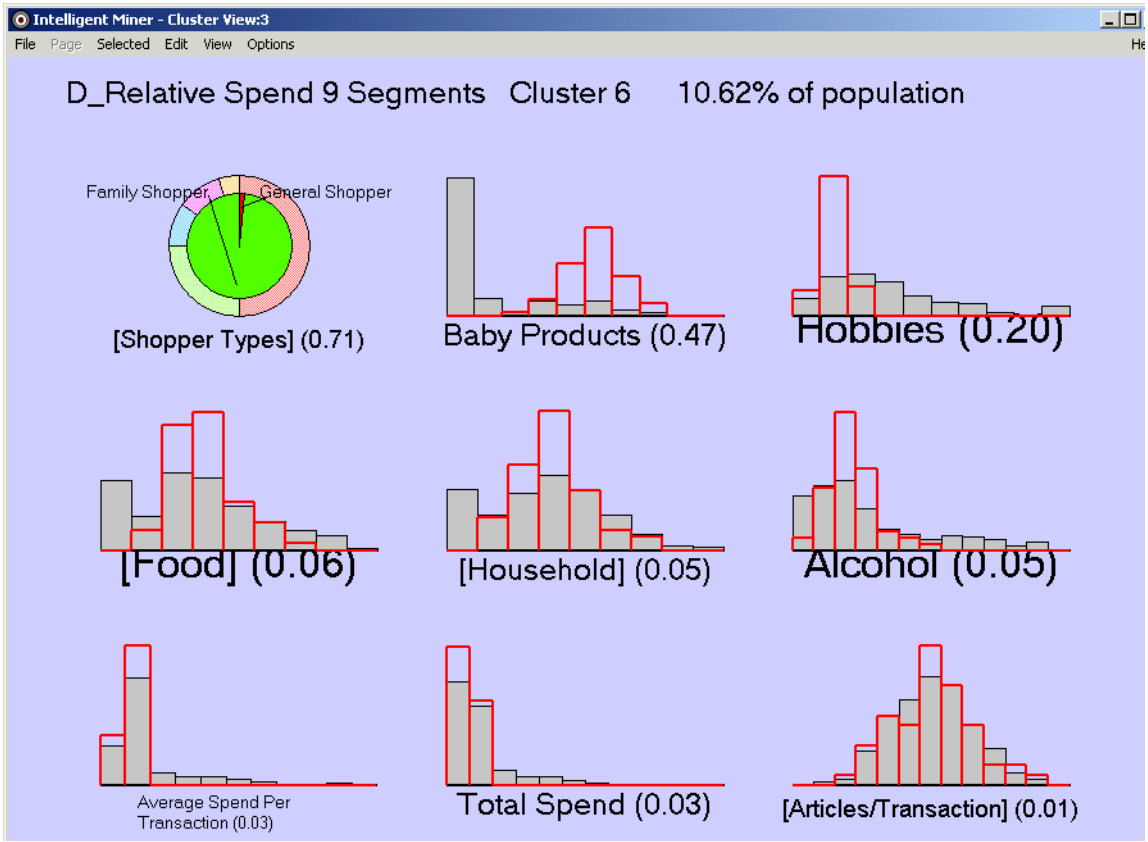


Figure 4-11 Demographic Cluster 6

To get a better picture of the differences between the two clusters, we can quantify what we concluded from a visual inspection by drilling into the detail and looking at the underlying statistics. These are summarized in Table 4-1 for the two clusters.

Table 4-1 Cluster 8 and Cluster 6 summarization

Characteristic	Cluster 8 Mean Value	Cluster 6 Mean Value	Ratio 8 to 6
Baby Product	4.5	5.4	120%
Hobbies	0.6	0.3	50%
Food	1.6	1.7	106%
Household	1.5	1.7	113%
Alcohol	0.4	0.7	175%
Total Spend	$2550	2350	92%

We can now see that although the Total Spend is about the same (a difference of 8%) the spend on Hobby products is 50% of that for the other cluster but this is compensated for by increased expenditure in Alcohol 175% and Baby Products 120%.

So although we thought we had only one group of Family Shopper, we really have two distinct groups of Family Shopper and this type of analysis helps us to identify why they are different. We could apply a similar analysis to the other clusters.

4.6.2 How do we compare different cluster results?

We explained in the previous section that different clustering techniques will produce different views of the same customer data. In our example we have generated two cluster results that on inspection appear to be similar. But how similar are they and how do we compare the two sets of results?

As an example we can see that both techniques produce two clusters that predominantly comprise Family Shopper types. We looked at the details for the demographic clusters above. If we now look at one of the two Family shopper clusters (Cluster 5) produced by the neural clustering technique, shown in Figure 4-12 we can see that this is not identical to either of the demographic clusters but seems to be a mixture of both.

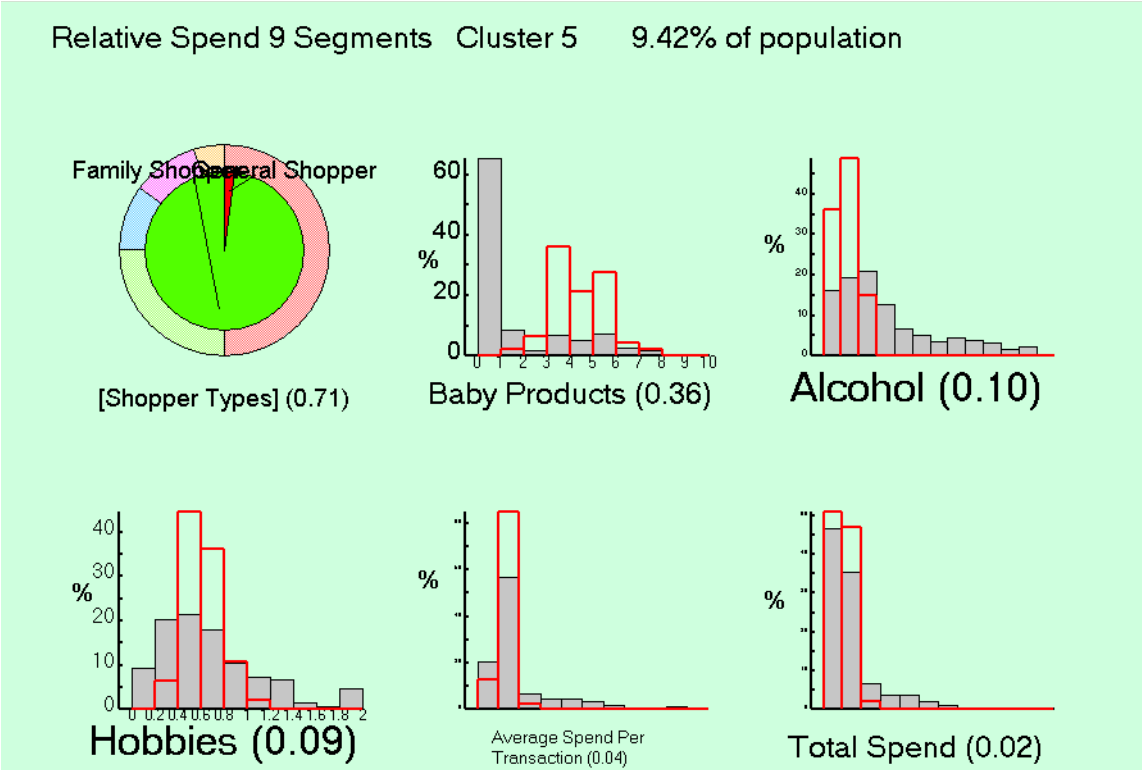


Figure 4-12 Corresponding neural cluster (Cluster 5)

Does this result mean that the analysis we performed above is incorrect and that there is a third way of looking at the customers in this category and what about the other clusters?

To find out how well our two cluster results compare, we can use a simple technique that makes use of the capability of our clustering techniques to create an output table that provides the details of which cluster a customer has been assigned to. We can combine the information from both output tables to create a display that allows us to make a visual comparison. An example of this is shown in Figure 4-13.

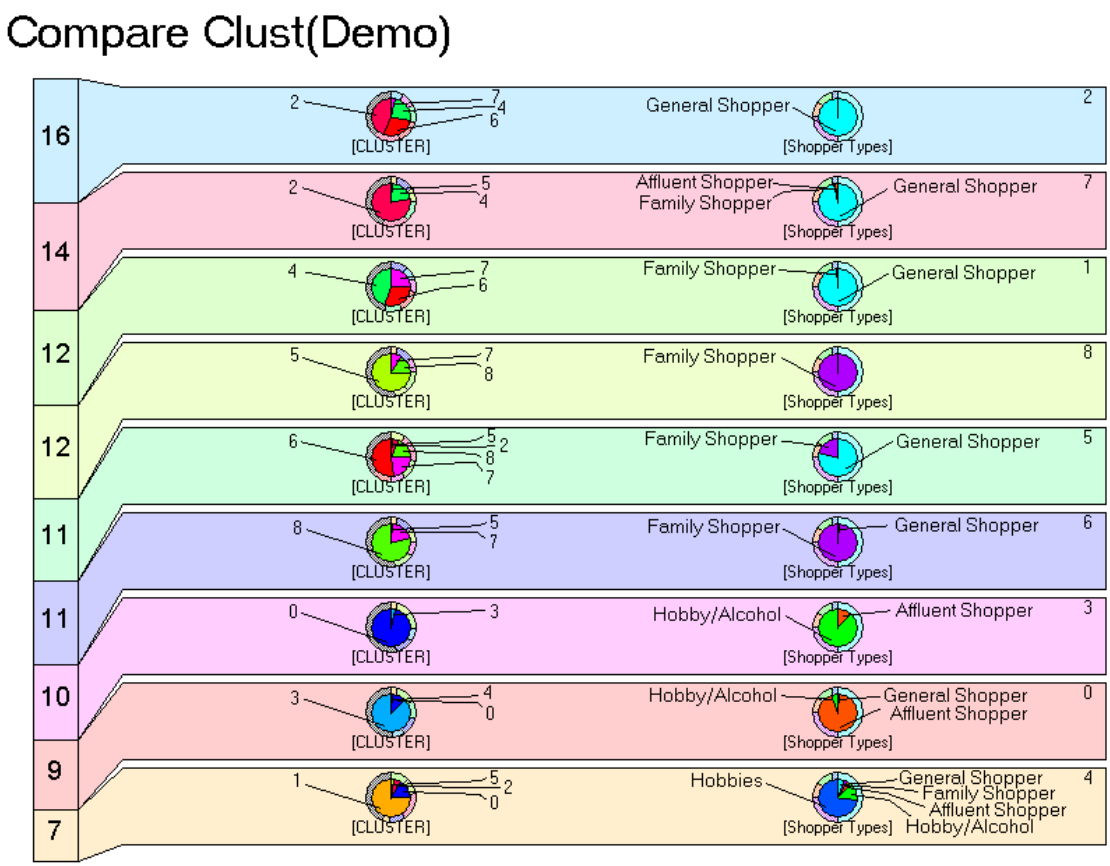


Figure 4-13 Comparison of neural and demographic clusters

In Figure 4-13, we show for each of the demographic clusters, the distribution of customers from the different Shopper Types within each cluster (as we have shown before) together with and the distribution which neural clusters these shoppers are also assigned. If we look at Cluster 6 and 8 we can see that these are a mixture of neural Cluster 8, 5 and 7, which confirms what we just discovered above. So which view is correct, or are both views incorrect?

To answer this question you need to understand in a little more detail just exactly what the two different clustering techniques are trying to do. Without going into all of the mathematics a simple way of understanding the neural clustering technique is as follows. The actual neural clustering technique we have used is grandly called a “Kohonen Self Organizing Feature Map”. What this means is that it attempts to group customers together as if they were on a checker board with

each square representing a cluster. In the case of our neural clustering technique we specified nine clusters and these are arranged in a 3-by-3 square, as shown in Figure 4-14, with Cluster 0 in the top left hand corner of the board and Cluster 8 at the bottom right hand corner.

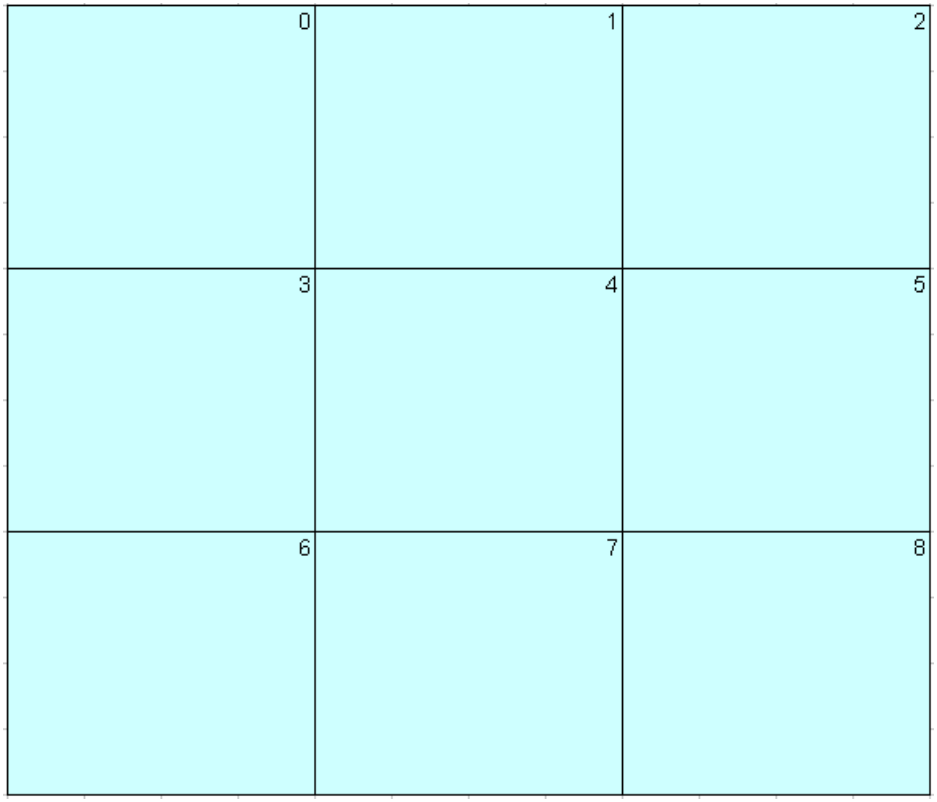


Figure 4-14 The neural cluster checker board for 9 segments

We can first try to draw where our business rule Shopper Types would be placed on this board by looking at Figure 4-9 on page 78. If we look at the top cluster (Cluster 2 we can see that this is entirely comprised of General Shopper customers, so we can shade in the square corresponding to Cluster 2 in the same way. The next cluster (Cluster 7) is predominantly General Shopper customers but with some Family Shoppers, so we can fill in the corresponding square appropriately. Doing this for all of our clusters will produce picture shown in Figure 4-15 where the numbers of the checker squares are actually cluster IDs from the segmentation.

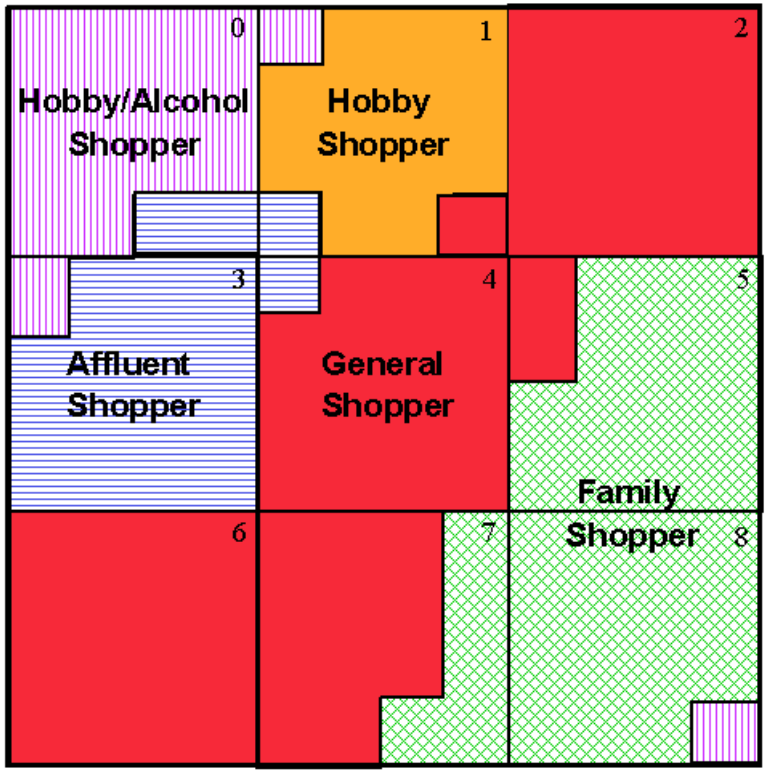


Figure 4-15 Distribution of Shopper Type on a 3-by-3 neural cluster checker board

What we can see immediately is that the neural clustering has done a pretty good job, not only has it put similar customer types together, but it has tried to keep the more dissimilar types apart.

Note: There is one interesting feature. This is a strange checker board in that customers who fall off one edge have a tendency to crawl underneath and appear on the opposite edge. You can see this in square 8 where some Hobby/Alcohol shoppers from Cluster 0 have slipped across. If your brain is big enough, try to imagine that instead of a flat checker board it was a rubber sheet that you could stretch and fold it into a sphere and you will see why this happens.

The neural cluster has a tendency to produce roughly similar sized clusters and because of this tendency, you can now see what happens if we increase the number of clusters for the neural algorithm. It simply creates more squares on the checker board and then moves the customers around to fit. An example of this is shown in Figure 4-16, where we have increased the number of neural clusters to 36 on 6-by-6 board.

Figure 4-16 Distribution of Shopper Type on a 6-by-6 neural cluster checker board

To produce this picture we have used a simple spreadsheet macro to shade in the squares, and where there is more than customer type per square we have just shaded it to correspond with the most predominant type. This means the boundaries between the different types may be slightly different from where they are drawn but not by much. Of course the neural clustering algorithm did not

know where to start its numbering and so the result is just rotated anti-clockwise through 90 degrees. In other words, what we originally called Cluster zero is now a combination of clusters 24, 25, 30 and 31. You can see this by just rotating the page clockwise through 90 degrees.

So what about our demographic cluster technique, where does it fit into this picture? Well you have to understand that it is looking at the world in a different way. It is trying to find groups that are most similar and is not trying desperately hard to produce even sized clusters. If we draw our demographic clusters on top of the 3-by-3 checker board as in Figure 4-17, we get a slightly more complex picture than in Figure 4-16, because we have nine categories rather than five shopper types, but we can now see how the demographic clusters fit in relationship to the neural clusters and in relationship to each other.

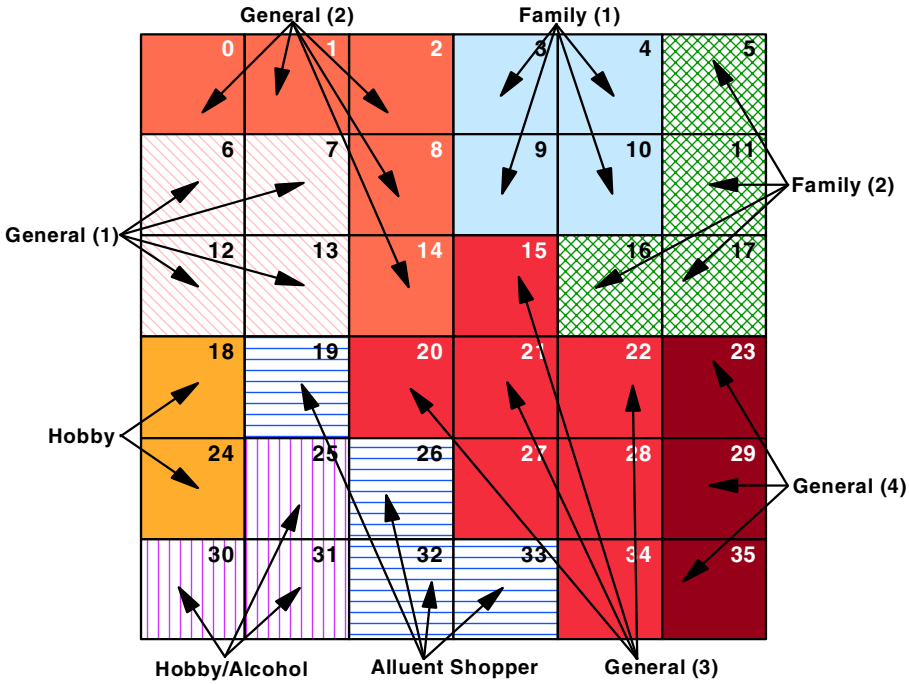


Figure 4-17 Demographic clusters imposed on a 6-by-6 neural cluster map

We can now see that in fact both clustering techniques have produced the same result but have presented their conclusions in different ways. Where the demographic clustering algorithm has tried to group similar customers into large segments, the neural clustering algorithm has distributed the same customers across a number of clusters.

4.6.3 What does it all mean? — Mapping out your business

The question we posed was which clustering technique should we use to describe our customers? The answer is that we use both. In fact we can now clearly see that we have a way of combining not only the two cluster results but also our original business rule segmentation into a single view of our business.

The results of course show us what we knew all along, that we don't really have distinct groups of customers but a continuum of different types of customer. However, we can now create through the power of data mining, a completely new way of visualizing our business. We can now picture not only where our customers are in relation to segments to which they have been assigned, but also in relation to each other and our other segments. It is also possible to position individual customers onto this map, and so we can see exactly where any specific customer is positioned.

We can also see that by increasing the number of neural clusters that we generate we have an easy way of increasing the resolution of our picture. The technique just gives us a better granularity as we increase the number of clusters. So why do we need to combine the neural and demographic cluster results? Where the neural clustering technique lays down the map the demographic clustering technique provides a way of putting the contours onto the map. Since demographic clustering uses the similarity between our customers to group them together in an appropriate way, this produces contours of similarity and gives us the scaling that we can compare our different groups of customers.

Continuing the map analogy by combining the two types of cluster result we can identify isolated hills on our map where we have areas of niche customer behavior that may require special attention in comparison to the large plains where we have more homogenous customer behavior. The areas between the different groups are also interesting, particularly where there are valleys and ridges of opportunity (for example, Clusters 20, 21, 25, and 26 in Figure 4-17). Customers in these regions are almost certainly the ideal candidates for cross-selling.

When we map out our customers in this way, we can immediately see the benefits of data mining to understand our customers better. It gives us a new view on what they are doing and this immediately prompts us to start asking more questions about what is going on.

Although we have not shown an example here, in the same way that you map your business rule segments onto the neural segments, you can do the same thing for other types of customer classification. As an example, you could segment your customers in terms of profitability or any other business measure, and then map this onto the neural clusters and see if there are regions where particular type of customer behavior gives rise to particular categories of the business metric. You will have to try this out for yourselves to see what we mean.

Data derived segments from transaction level data

In this section we have concentrated on segmenting customers who we can identify using loyalty card information and for whom we can aggregate transactions to produce the CLA data model. The main reason for doing this was to show how to map the predefined business rule segments onto the data derived segments. If you only have point of sale transaction level data, but no means of performing customer level aggregation, then can you still perform data derived segmentation?

If you perform clustering using the TLA model, then you will also produce a segmentation that reflects the characteristics of your customers in terms of the combinations of products they purchased during a single transaction. In this case the picture is somewhat different from the CLA model results, as shown in Figure 4-18.

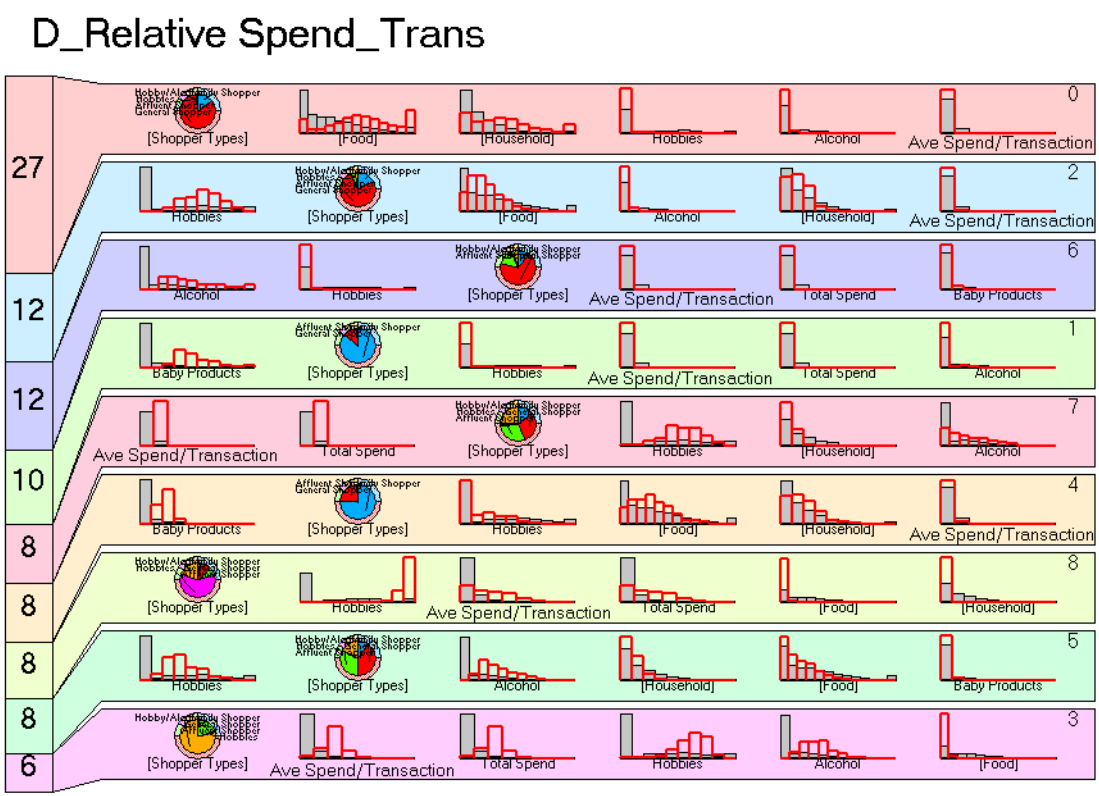


Figure 4-18 Segmentation using transaction level data only from the TLA model

In Figure 4-18 we have included the Shopper Type variable for comparison with our previous segmentation results. If you perform this type of segmentation and you do not have loyalty card customers, then you will usually not have the shopper type label and you will have interpret the segments and associate a business meaning to them. However, as you can see in Figure 4-18, although the segments are now a mixture of our original shopper types, there are still segments that are mainly what we defined as General Shoppers (Clusters 9, 2, and 6) and Family Type Shoppers (Clusters 1 and 4) and Cluster 3 is still predominantly Hobby shoppers. So in this case you would still define similar types of labels to your different groups of customers. Once you have assessed the segments and labeled them appropriately, then you are again in a position to develop your business strategies around the segments and everything that we have discussed in the previous sections applies.

4.7 Deploying the mining results

The final and *seventh stage in our generic mining method* is perhaps the most important of all. How do you deploy the mining results into your business and derive the business benefits that data mining offers?

The reason that deployment is so important, is that all to often data mining is seen only as an analytical tool and not as the means to drive your CRM systems. In this section we explain how the data derived customer segments can be deployed into your business. We specifically address how recent advances in data mining technology enable you not only to export results but also the clustering models that you create and how these can be imported and used by CRM and other tools within you business.

4.7.1 Scoring your customers

The clustering techniques that we described in the preceding sections produce two types of output.

The first type of output are the cluster results themselves. In this type of output each customer record can be scored in terms of the following variables:

- The cluster/segment to which the customer has been assigned
- The affinity of the customer record to the cluster expressed as a cluster score
- The next nearest cluster/segment
- The affinity of the customer to this segment
- A measure of the confidence that the customer belongs to the cluster/segment to which they have been assigned

The second type of output is a cluster model, which can be used to score customer records that were not used to generate the clusters. The cluster model can be stored, and reused by the data mining tool itself to score customer records held in other databases, or files to which the mining tool has access (application mode). The alternative and potentially more valuable method, is to export the cluster model in a format that can enable the models to be integrated into other applications. This format is known as Predictive Model Markup Language (PMML) which provides a standard by which models from different mining tools can be exchanged. Exporting models in this way enables other applications, for example relational databases such as DB2 or Oracle, to import the models and use them within the application to score customer records directly and when required by the application. This is an exciting development for the retail business sector, since it now enables you to deploy the results of your data mining into a wide range of applications (for example, point of sales systems, kiosks, campaign management tools, Web based services).

In the following sections we look at a couple of examples of how customer scoring can be used.

4.7.2 Using the cluster results to score all your customers

The usual application for deploying the cluster/segmentation results is where, following an analysis of the data derived segments, you want to be able to classify your customers in particular segments and use this information in some specific way. A number of campaign management tools (for example, Xchange Campaign) require you to provide a table of customers and the segments to which they have been assigned. Alternatively, you may want to use the segments as additional dimensions in some type of On-Line Analytic Processing (OLAP) tool and perform some more detailed investigation of the customer attributes. In this case a similar table is required.

The simplest way to generate this type of table is to use the first of the two methods described above. This method assumes that the data mining tool has direct access to the appropriate customer records and that these records are stored either as a single table, or an equivalent database “view”, at which the data mining tool can be directed. Using the application mode of the data mining tool then enables all customers to be scored and assigned to a segment with the results being written back the another database table or file. This is illustrated in Figure 4-19.

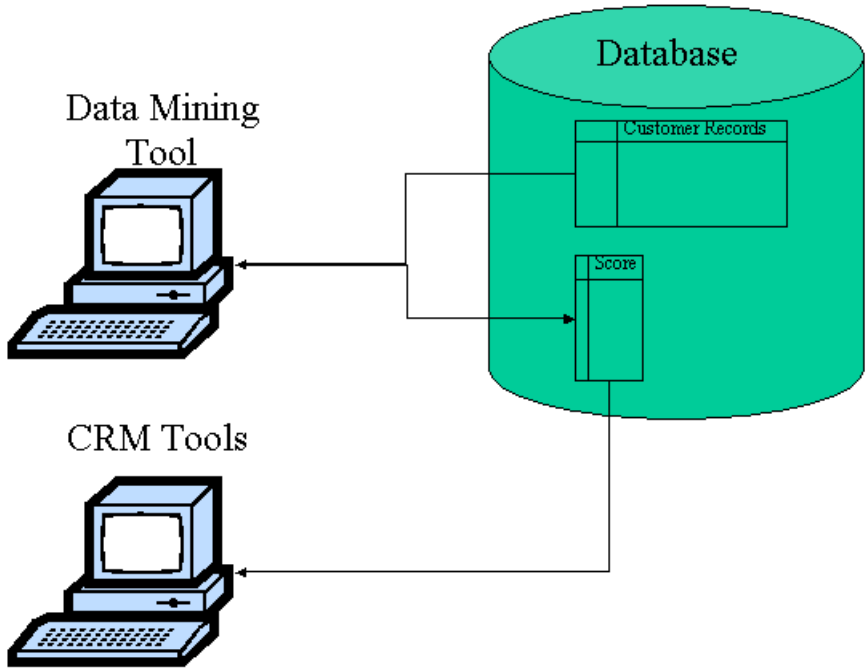


Figure 4-19 Scoring using the data mining tool directly

There may be situations where it is not possible to perform this operation directly from the data mining tool or where scoring all of your customers is either unnecessary or desirable. For example, the clusters may have been developed using a subset of customer records held in a separate database from the one containing the customer records you want to score. In such cases the use of the second method of scoring may be more appropriate. This is discussed below.

4.7.3 Using the cluster results to score selected customers

While scoring all of your customers can be performed as some form of batch process, if you have large numbers of customers this may introduce a significant overhead on database performance. If you have to score all your customers this may be unavoidable, but in many situations there may only be a subset of the customers that it is appropriate to score at any one time. This is increasingly true for e-commerce applications where you may only want to score customers as they use a particular service. Using the exported PMML models, it is possible to score selected customer records directly within the database. This type of scoring is performed using available database functions and therefore can be done automatically, for example, as customer records are being updated or at

specified times. An additional advantage of this approach is that the PMML models themselves are stored in the database, and therefore the configuration control of which models were used and when, can be handled by standard database administration procedures. The process is illustrated in Figure 4-20.

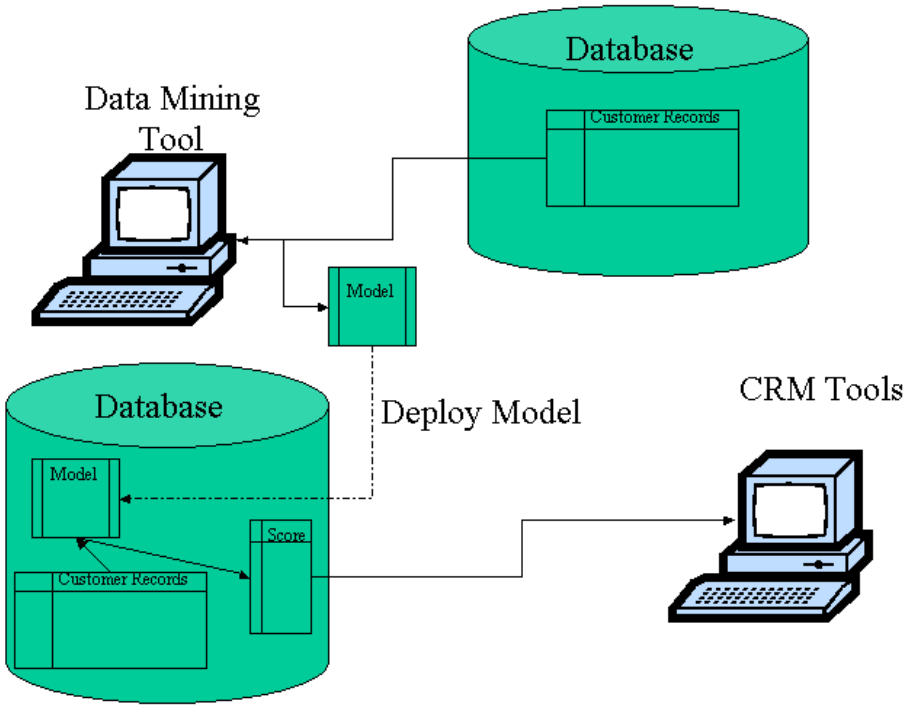


Figure 4-20 Scoring customers using exported PMML models

Using this type of approach for deploying data mining solutions into your business offers many advantages. It makes possible the deployment of the same models into distributed applications being used across your business. The same models, although developed by your operational research department, could be being used by your sales and marketing organization (in OLAP and other forms of reporting tool such as Business Objects, BRIO and so on), at the same time the models can be deployed into your CRM systems (kiosks, Web-shopping and similar customer touch points) and as part of your campaign management tools (direct mailing, telesales and so on). As the models are updated or improved, they can easily be replaced in all of the tools that you are using and consistency is then maintained.

There are an increasing number of applications that can make use of the capability to deploy data mining models in this way and as e-commerce expands in scope, there will be many more opportunities to exploit this capability. In Chapter 6, “How can I decide which products to recommend to my customers?” on page 137, we give another example of how this capability can be used as part of a personalized product recommendation system.

What needs to be stressed is that the deployment of the results of data mining into your business can bring significant benefits and wide ranging benefits. It is worth spending time considering how you are going to use the results before embarking on any data mining activity.

5

How can I categorize my customers and identify new potential customers?

In the previous chapter we looked at ways of discovering market segments using the data you routinely collect about your customers. There are however many situations where, rather than trying to discover new segments, you may want to classify your customers into existing segments or other predefined categories. Typically, this type of requirement is needed where you have used specialized customer information to define the categories but where the information was only available for a small subset of customers (for example, customer survey data or loyalty card information). You now want classify all of you customers into these categories. The question is, can you do this using the data you routinely collect on all your customers?

Alternatively, you may have already categorized your customers using routinely collected data and now have some additional data that can be linked to existing customers and to potential new customers (for example, demographic data). Can you discover potential new profitable customers by using the demographic data alone?

In this chapter we look at the data mining techniques that enable you to answer these types of questions.

5.1 The business issue

In the previous chapter we looked at the issue of how to derive customer segments from the data you routinely collect. Using the data mining technique of clustering we saw how it was possible to derive customer segments without any prior knowledge of the different types of customers that you have. This was an example of what we called *discovery data mining* in 3.3.1, “Types of techniques” on page 27.

We also had available an existing business rule segmentation and we were able to show how these business rule, derived segments could be mapped onto the clusters we had discovered. To do this mapping effectively required customer transaction data to be aggregated over a number of transactions and this in turn required some means of linking transactions through a customer identification number. An obvious question to ask is can we use another method to match customers to pre-defined categories like our business rule segments and if so could this be applied to single transaction data? The answer is yes, and the data mining technique that we use to do this is called classification and comes under the general heading of *predictive data mining.*

There are many potential applications for classification within your retail organization. In fact in almost any situation where you have categorized groups of customers (or indeed anything else you can think of), you can use classification to discover how the associated data can be used to classify other customers into the same categories.

One example of this process is where an existing customer segmentation has been derived using some specialized information, obtained, for example, from customer surveys or focus groups. This type of information is often expensive to collect and you may only have it available for a small subset of you customers. You will also have available for this subset of customers the routinely collected data that you have for all your customers (for example, transaction data). If this data can be used to determine to which segment a customer in the subset belongs, then clearly you can use the same data to categorize all of your customers.

Alternatively, you may have derived your customer segments using routinely collected data, as we did in the previous chapter, but now you want to identify potential new customers that could be matched to the most profitable segments. In this case, if you can obtain external data (for example, demographic data) that can be linked to both existing and potential new customers, then you can perform classification with the demographic data to identify the segments that existing customers belong to and then use this to identify potential new customers that would fall into the most profitable segments?

As we will show, classification can be used to do all of these things but in general it is not possible to achieve a 100% correct classification for all of your customers. The issue then becomes a question of how confident do you need to be in identifying customers to the segments in order to use the classification results effectively. In the following sections we show you how to use classification to categorize your customers, how to interpret classifier performance, and most importantly provide some suggestions on how to deploy the results into your business, either as part of a directed marketing campaign, through point of sale systems, or other customer touch points.

5.1.1 Outline of the solution

The *first stage in our generic data mining method* is again the translation of the business issue into a set of questions that can be addressed by data mining. In the case of classification, there are a number of data mining techniques that can potentially be used. The challenge is to identify the most appropriate technique for the business issue that is being addressed.

All of the data mining classification techniques construct a mathematical representation of your data that relates the different characteristic variables of your customers to the predefined categories that you have assigned to a subset of your customers. We call this mathematical representation a *classification model*. The example we will describe in this chapter demonstrates how to construct such a model and then how it can be used to classify the customers that have not been pre classified. The predefined categories can be anything you choose, for example, the customer segments we have already described, or categories of profitability, or even whether your customers are likely to move to a competitor.

To illustrate how to perform classification we are going to use the predefined business rule segments for the loyalty card customers defined in the previous chapter. We already know that if we use the aggregated NRS, then these customers can be successfully mapped onto the business rule segments. The question that we want to address is, Can we use the data from single point of sale transactions to do the same thing? If we can, then because we have this information for all of our customers and not just our loyalty card customers, we can use the classification model to categorize all of our customers at the point of sale and make appropriate offers to them. The data mining techniques that can be used to determine what types of offers to make are the subject of Chapter 6, “How can I decide which products to recommend to my customers?” on page 137.

To build a classification model it is first necessary to have data on a group of customers that have already been assigned to business segments. We call this group of customers the *training group*. The data that we have on these customers is divided into a *training data set*, that we use to develop the classification model, and a *test data set*, that we use to validate the model. In our example, the data required are the individual transaction records, but this could just as easily be some demographic data or other information that you can obtain for both the training group and the group of customers you want to classify. We call this second group of customers the *target group*. The data that we have for the target group is of the same type as the training group, but for these customers we do not have the predefined business categories. We call this type of data *operational data*, since it is the type data that we want to use when we deploy our classification models into our business.

The concept behind classification is that we construct a classification model using a training group of customers, and then use this model to classify customers in the target group. This is illustrated in Figure 5-1.

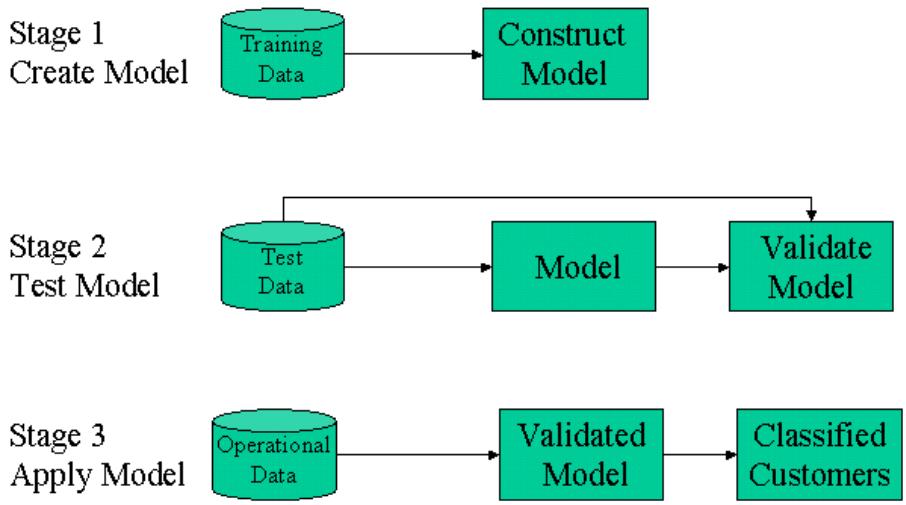


Figure 5-1 Training and testing and applying a classification model

5.2 The data to be used

Stage two of our generic mining method is to identify the data that is to be used to construct the classification models. In our example, we use both the CLA data model and TLA data model described in 4.2.2, “Suggested data models” on page 53, and use these data models to construct classification models that can be used to categorize customers into the predefined business rule segments.

Because our training group customers were originally assigned to business segments using the aggregated NRS from the CLA model, we first use our data mining classification techniques to show how well the data supports the classification that was originally made. We then use the TLA model to show how well our target customers can be classified to the same business segments when we restrict ourselves to using the NRS from single transactions.

To create the classification models it is necessary to extend both the CLA and TLA data models. In both cases, each customer record in the training group must include the business segment label which will then become the target variable for the classification. It is also necessary to create some additional variables that will be used by some of our classification models. These additional variables are derived from the business segmentation, with one additional variable for each business segment category (for example, General Shopper, Family Shopper variables). The value of each new variable is set to “1” if the business segment label matches the variable name, and is set to “0” otherwise. When we come to discuss the construction of our classification model, we use the term *target variable* for these new variables. The structure of the modified CLA and TLA data model is shown in Figure 5-2.

Customer_ID	NRS_FOOD	NRS_ALCOHOL	NRS_.....	Shopper_Type	General_Shopper	Family_Shopper	.....
12345	1.2	0.8	...	General Shopper	1	0	0
23451	2.5	0.5	...	Family Shopper	0	1	0
31245	0.1	0.1	...	Hobby Shopper	0	0	0
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Figure 5-2 Structure of the CLA and TLA data model table

5.3 Sourcing and preprocessing the data

If you have already sourced the data for doing segmentation then you will not require any additional information to perform this *third stage of the generic mining method*. However, as we have already discussed, to construct a classification model it is necessary to create two separate samples of your training group customer data, the training data set and test data set. To do this requires some specific preprocessing steps to be taken.

5.3.1 Creating the training and test data sets

The training set is used to develop the initial model. After the model has been developed the test data set is used to validate the model by measuring its performance on what is usually termed *unseen data.* A key part of validation is to check for a phenomenon known as *over fitting.* Over fitting occurs when a model is developed to such an extent that it only fits the training data, but cannot be

generalized to other data sets. To avoid this, the training and test data sets must be both independent and wholly representative samples of the training group of customers. This requires that the data in the training and test data set must be randomly selected from the database table or file containing the TLA data model data and placed into one or the other data set.

Stratified sampling

In the case where you have a large number of customers and large variations in customer numbers in the different segments, it is sometimes necessary to developing test and training data sets by sampling from all of the data that you have available. In these cases, it is important that the customer records are sampled proportionately from each segment or category. This technique is called *stratified sampling*. In some cases, although a single segment or category name has been assigned to a group of customers, there still may be significant differences in the characteristics of the customers in that segment. We have already seen this in the case of the General Shopper segment in Chapter 4. In these cases the sampling can be improved by first clustering all of the customers from one category at a time and then sampling from each cluster. This is an advanced approach to stratified sampling that has been shown to produce significant improvements in classification performance where the customer segments are not homogeneous.

Note: One of the advantages of data mining products such as IM for Data is that you can mine all of the data and in these cases stratified sampling is not required.

Balanced samples

There is a general misconception, so that it is necessary to create what is known as a *balanced sample* before classification models can be developed. The misconception stems from the idea that if there is a large variation in the number of customer records in a particular category or segment, then it is necessary to compensate for this by creating a sample that contains the same number of customers from each category or segment. As we will show, classifier models produce an output that estimates the probability that a customer belongs to a particular category. If the training group is a representative sample of your customers and you do not have any reason to favour one segment or category above another, then there is no reason to create a balanced sample.

You can understand why this should be the case by considering a simple example. Imagine that all your customers had identical characteristics in terms of their NRS, but that you had allocated them to segments or categories using some other information. If you try to classify these customers using the NRS data, then one customer is clearly indistinguishable from another. The probability

that they belong to a particular segment is then the ratio of how many customers were originally assigned to that segment compared to the total number of customers. (If 90% of customers are General Shoppers, then, if you have nothing else to go on, the probability that a customer selected at random is a General Shopper is still 90%.) If you had used a balanced sample the classifier would have concluded that the probability was equal for all segments (inversely proportional to the number of segments) which is not the case.

You should normally only use a balanced sample where you know that the training group is itself an unbalanced representation of the segments or categories within the total customer population. For example, you may know that in the total customer population there are equal numbers of customers in each category, but in your training group there is an uneven distribution due to some sampling reason. In this case, balanced sampling can be used to redress the balance. The other situations in which balanced may be appropriate, is where the risk of incorrectly classifying customers into one category, rather than another, is important to you. This can also be achieved using a process called *error weighting* and we will have more to say about this issue when we look at the different types of classifiers in 5.5, “The mining technique” on page 104.

In the case of our example data set, we are able to mine all of the data and so we do not require any stratified sampling. We also know that the training group is a representative sample of the total customer population, and since at this stage we have no reason to favour one group of customers over another, we do not require any form of balanced sample. Therefore, the test and training set can be developed by randomly splitting the data, with 50% of the training group customers in the training data set and the other 50% in the test data set.

The important matters to stress about the preprocessing step are that you need to think carefully about:

- How representative the training group of the customers is, in comparison to those customers you going to classify using the resulting model.
- How you are going to use the classification results, and whether one category is more important to you than another.

5.4 Evaluating the data

Evaluation of data includes resolving problems with missing values, outliers and redundant variables. This is *the fourth stage in our generic mining method,* and because we are using the same data models as those in Chapter 4, then the same procedures as those covered in 4.4, “Evaluating the data” on page 63 should be followed.

Most classification models are very sensitive to highly correlated characteristic variables, and therefore the evaluation steps that we described in 4.4, “Evaluating the data” on page 63, to remove or combine such variables need to be considered depending on the type of classification technique you are going to use. Some of the reasons for this are discussed in the following section.

An important part of the evaluation stage is to ensure that the test and training data sets accurately reflect the statistical characteristics of the total customer data set. A good way of validating that you have split your data in an appropriate way is to look at the statistics for these data sets (mean, standard deviation and modal values) using the univariate statistics and check that the statistics of the test and training sets still match the statistics of the whole data set.

5.5 The mining technique

The *fifth stage in our generic data mining method* is to identify and choose the appropriate data mining techniques that we are going to use and to determine how we are going to apply them to the specific business issue. In the case of customer classification there are a number of different data mining techniques that can be used. To decide which technique is most appropriate, you need to understand how the different techniques construct their classifier models and how to interpret them. In this section, we will look at some of the techniques you can use and how they are applied.

5.5.1 The classification of mining techniques

It is usual when you are performing customer classification to use more than one technique and then to compare or combine the results to achieve the best overall classification performance. The reason for doing this is that the different techniques perform their classification task in different ways and, just like the clustering techniques in the previous chapter, this leads to variations in performance. Among the most popular techniques are:

- Decision trees
- Neural networks
- Radial Basis Functions (RBF)

Here we use two contrasting techniques, the decision tree and RBF classifiers. The main advantages of the decision tree technique is that it delivers good performance, it can perform multiple category classification and has the most accessible results that are relatively easy to interpret. In contrast the RBF technique usually offers better performance, but the results are less easy to interpret and classification can only be performed for one category at a time.

The reason for focusing on the two different approaches is to illustrate the differences, to describe the steps you should go through when developing classification models, and to show how the results from two classification techniques can be combined to increase your confident in the final results.

5.5.2 Decision tree classifiers

A *decision tree* classifier is constructed from the training data set of customer records by progressively splitting the customers into smaller groups. The splitting is performed such that each new group is a purer sample for one of the customer categories than the original larger group. For example, if the original group comprised a mix of two types of customer, say General Shopper and Family Shopper, then if the group could be split into two groups with predominantly General Shoppers in one group and Family Shoppers in the other, then the desired objective will have been achieved. A measure of the purity is the ratio of the number of customers in the predominant category for the group divided by the total number in the group. If by splitting the group, the average value of this measure increases, then the split has improved our ability to identify the different categories of customer. For example, if we select a customer randomly from one of the groups and label it as being in the predominant category, the chance of being correct is now higher.

Note: The actual measure we use to determine the quality of the split is called the GINI index. The GINI index measures the purity of the split and then weights this by the number of customers in each group. This results in a split that minimizes the error, while avoiding trivial splits where just a few customers of one category are separated from the rest.

The decision of how to make each split is made by looking at each characteristic variable and finding the variable and the value of that variable that leads to the purest split. Once a split has been made, then each group resulting from the split can itself be split again using other variables or even the same variable. The process then continues, progressively splitting into smaller and smaller groups until either the group contains only one category of customer, or until an acceptable purity is achieved. This results in the type of tree structure is shown in Figure 5-3.

Note: This decision tree shown in Figure 5-3 has been generated from synthetic data to illustrate how the different classifiers work and not from the example data set that we will evaluate in section 5.6, “Interpreting the results” on page 118. The synthetic data set is comprised of two categories of customers, each with two contrasting characteristic behaviors, in their purchase of Baby Products and Food. In this case, General Shoppers are characterized either by relatively high spending in both Baby Products and Food, or by relatively low spending in both product categories. In contrast Family Shoppers are characterized by either high spending in Baby Products and at the same time low spending on Food, or conversely high spending on Food, when there is low spending on Baby Products.

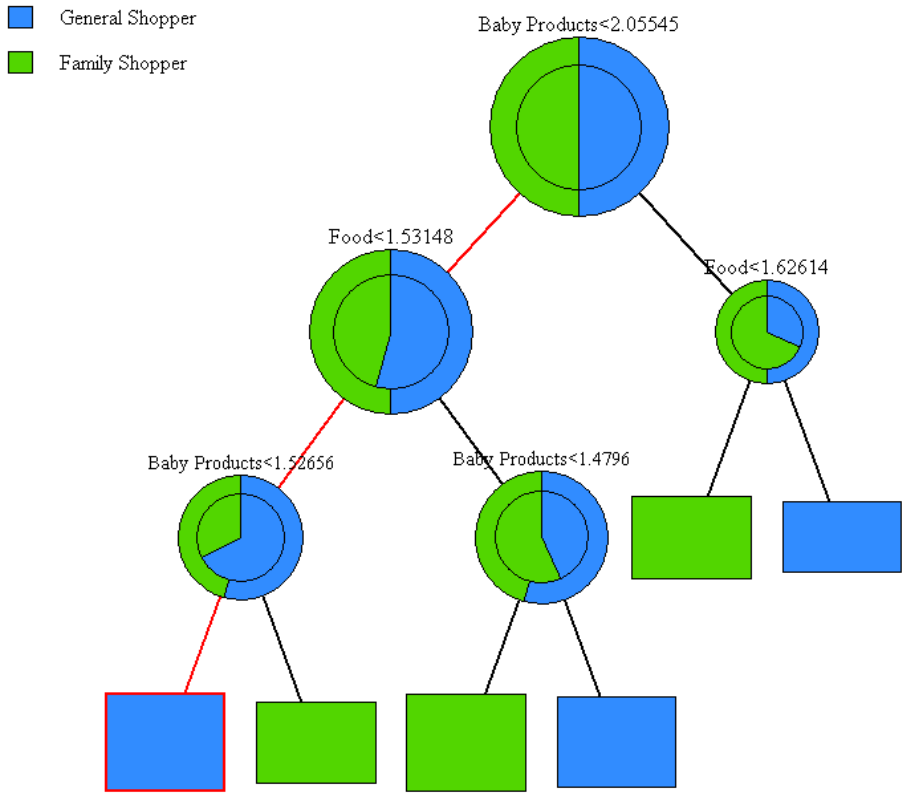


Figure 5-3 A simple decision tree

Although we call this a decision tree classifier, the diagram data miners always draw to represent it is of an upside down tree. At the top of the diagram, the first split is called the *root node* and continuing the tree analogy the splits result in *branches* and the nodes at the end of the branches are called *leaf nodes* or simple *leaves*. This does not imply that data miners go around with their heads in the sand, but we must say something about the convoluted way some people look at the world!

In general, the splitting can continue until either all of the leaf nodes contain only one customer type, or until an acceptable number of errors occur. In the limit the splitting can result in one customer at each leaf node in the tree. This is a good example of a tree that over fits the data, because although it would achieve 100% correct classification on the training data, it is very unlikely to produce the same result on the test data set. To prevent over fitting and to produce a tree that can be applied to unseen cases, the basic tree must be pruned to a level where an acceptable performance is achieved on both the training and the test data sets. This can be done in a number of ways from simple manual pruning to fully automatic pruning of the tree. In the latter case, the pruning usually attempts to balance the number of errors with the complexity of the tree. A very complex tree (many branches) will have fewer errors, but is less likely to perform less well on unseen data than a simpler tree with more errors. Because at each node the split is made by selecting one variable and an appropriate value of that variable, this type of tree is called a *binary tree*.

Note: The pruning technique is often referred to as *minimum description length* pruning and this is based on the same principle as Occam's Razor, which essentially says that things should be expressed as simply as possible but no simpler.

When you want to classify a new customer using the tree, the variable for the target customer is compared with the value of the variable at the top of the tree and the appropriate branch is followed depending on the outcome. This continues until a leaf node is reached where the customer is classified according to the distribution of records from the training set at the leaf node. A confidence of being in the class is then calculated. The path through the tree can be expressed in the form of a rule, for example:

```
If Baby Products > 2.1 (relatively high spend)
and Food < 1.6 (relatively low spend)
THEN
Class = Family Shopper with a confidence of 100%
```

This shows that the tree has correctly identified one of the Family Shopper groups we defined and makes it relatively easy to understand how the classification decision tree has been made. This transparency is one of the main advantages of the decision tree technique.

Variable selection and preprocessing requirements

The explicit selection of one characteristic variable at a time when making each split is what makes the decision tree easy to interpret. Since successive splits can be described by the type of rule illustrated above. At the same time, this type of feature selection is the main limiting factor to achieving optimum classification performance. You can understand the reason for this by considering the situation illustrated in Figure 5-4 which shows the distribution of two categories of hypothetical customer (A and B) plotted against two characteristic variables (V1 and V2).

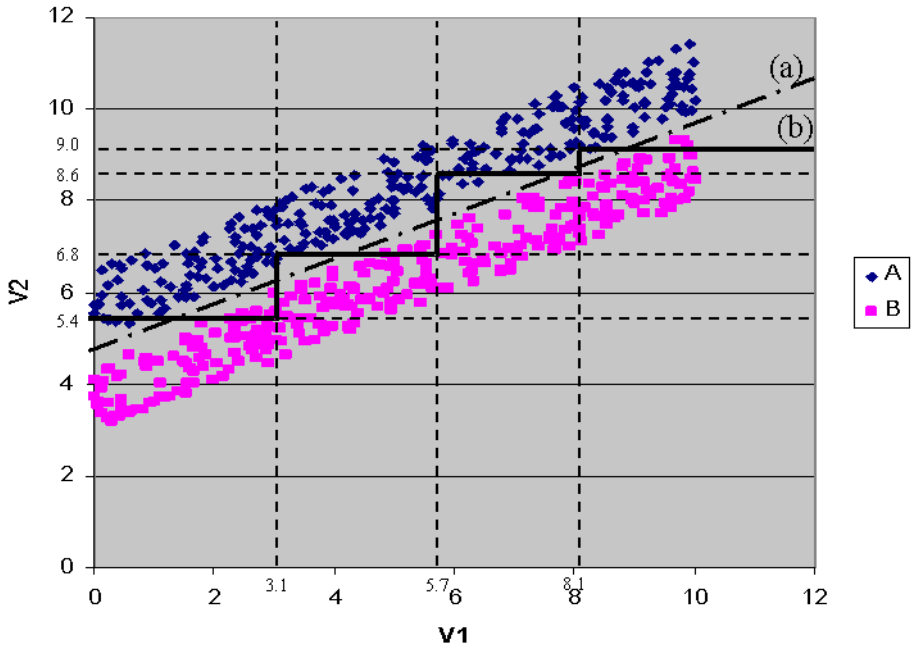


Figure 5-4 How a decision tree performs classification

In this case the two types of customers both exhibit a high degree of correlation between the two variables, V1 and V2, and can “best” be separated by using the dividing line (a) which bisects the two groups. If however we are restricted to using only one variable at a time to divide the two categories, then this cannot be achieved using a single split and in this case several splits using each of the two variables has to be made (at values 3.1, 5.7 and 8.1 for variable V1 and at values

5.4, 6.8, 8.6 and 9.0 for variable V2). The resulting boundary between the two regions is now the curve (b) which is the decision tree's approximation to the line (a). The order in which the splits are made can be understood from the corresponding decision tree is shown in Figure 5-5.

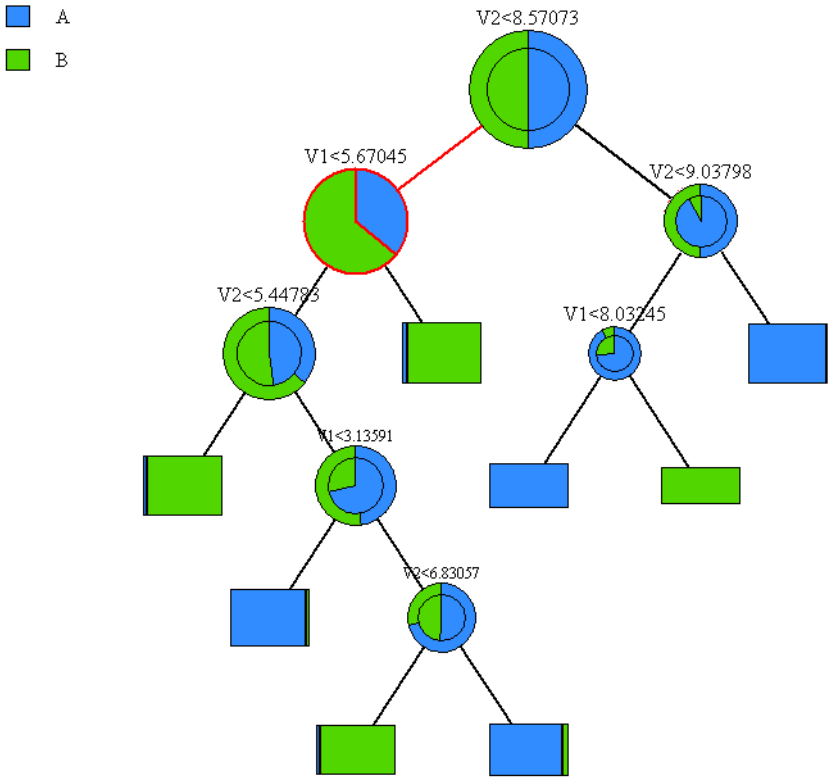


Figure 5-5 Decision tree corresponding to Figure 5-4

The decision tree can be thought of as describing the regions indicated by the dotted lines in Figure 5-4, but the net effect of applying the tree is the decision boundary (b). If it had been possible during the preprocessing step to calculate some new characteristic variable such that the line (a) in Figure 5-4 was perpendicular to the new variable, then only a single split would have been required and we would have a simple decision tree. To find such a variable's equivalent, do a rotation of the coordinates of the graph shown in Figure 5-5.

Note: Mathematically, this is just a weighted combination of the variables V1 and V2 where the weights are proportional to the gradient of the line (a). Although this is relatively easy to see in this case, in general with many variables, this is much more difficult to identify. The statistical techniques of Principal Component Analysis and Factor Analysis mentioned in 3.3, “Data mining techniques” on page 27 can be used to perform this type of variable transformation.

If the coordinate rotation had been done in this case, then we would end up with fewer splits and potentially a decision tree that had better performance on unseen data. The penalty that would have been paid is that the variables used to make the split would now be more complex, making the tree itself more difficult to interpret. Generally, there is always a balance to be struck between classification performance and your ability to interpret the reason for the classification decision, and just goes to prove the adage, “there is no such thing as a free lunch”.

Error weighting

When constructing any type of classifier you need to ask yourself the question: Are all classes equally important in my business decision process? Suppose for example, that you wanted to construct a classifier to classify General Shoppers and Family Shoppers. In this case, you may know that although a Family Shopper would not object to being labelled as a General Shopper, the converse is not the case (offering General Shoppers diapers might be seen as offensive, while offering Family Shoppers household goods is acceptable). You therefore need to build a classifier that can take into account the risk of incorrectly classifying a customer into the wrong class. The process that we use to perform this is called error weighting.

In this example, you want to weight the decision tree so that where there was a chance of making a mistake in the classification decision, this would be biased towards classifying Family Shoppers as General Shoppers, rather than the other way around. This can be done by biasing the split decision at each node in the tree in favor of the General Shopper class using a risk or error weighting. The challenge is to find an appropriate weighting that maximizes the opportunities to identify the target customer group while minimizing the incorrect or false classifications.

This type of error weighting can also be used where the class that is important represents a relatively small group of customers who can easily be confused with another class. This could represent a highly profitable niche group of customers within the overall General Shopper group. In this case by using an appropriate error weight for the training set would result in this group being identified, but with some general shoppers being classified incorrectly.

An alternative to error weighting is to use balanced sampling to create the training and test data sets. In 5.3.1, “Creating the training and test data sets” on page 101, we looked at the issue of creating a balanced samples. The equivalent to error weighting is to create a balanced, by over sampling from the categories that you want to boost. This is simply a question of using the same customer record multiple times in the training data set, but using it only once in the test data set.

Tip: Creating such samples for a multi category classifier, such as the decision tree classifier, is quite complex and in general it is better to use error weighting to achieve the result. In the case of a classifier that makes a binary decision, then over sampling is relatively easy to perform.

5.5.3 Radial Basis Function (RBF)

The RBF is a different type of technique that we can use to predict the category to which a customer should be assigned. Usually the RBF technique is used to predict a continuous variable (the target variable) that is a function of one or many other variables. If we want to use the RBF technique to perform classification, we restrict the target variable to be either “1” or “0”, and the RBF then predicts the category with a value in the range 0 to 1. This prediction is the probability that the customer is a member of the target category.

To prepare our data for use with the RBF classifier we first label each customer in our training set to be either a member, or not a member of the target class. We then specify a new binary variable. This was done in preparing the modified CLA and TLA data models that we described in 5.2, “The data to be used” on page 100. While the tree classifier can cope with multiple target categories, the RBF treats them one at a time and a separate classifier has to be constructed for each category.

Note: We could have used this same approach with the decision tree classifier. In some cases this can result in an improved classification performance when there is a specific category that we want to concentrate on.

The way in which the RBF constructs its model has some similarities with the decision tree classifier. The RBF divides the two categories of customers into regions using the characteristic variables to define the boundaries between the regions. The RBF technique is not limited to selecting one variable at a time, and therefore the boundaries between the regions can be lines (two variables), planes (when you have three variables) or what are known as hyperplanes when you have more than three variables.

The regions separate customers where the target variable has similar values (in this case “1” for Family Shoppers and “0” for all other shopper types, which in the case of this data set are just the General Shoppers). In each of these regions the technique places what is called a fitting center and at each center it places a basis function. The basis function measures the confidence that a customer belongs to the region and is defined so that the further away a customer record is from the center the lower is the confidence. Therefore, the name is Radial Basis Function for this type of classifier.

If we develop an RBF classifier model using the same synthetic data set that was used describe how the decision tree classifier builds its model, then the RBF generates four regions and four fitting centers as shown in Figure 5-6.

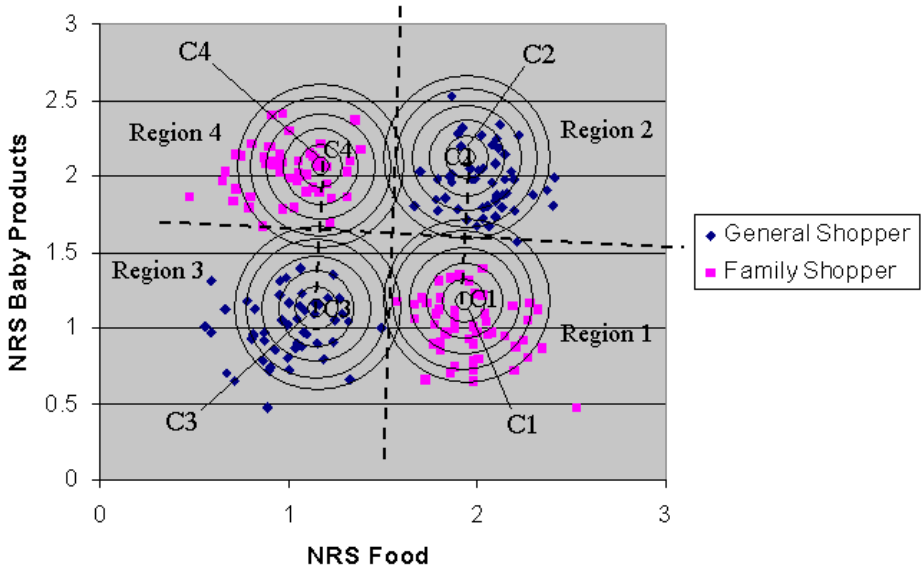


Figure 5-6 The Radial Basis Function

Because the splitting is not limited to using one variable at a time, this results in the region boundaries indicated by the dotted lines that are now not perpendicular to the variables on the graph axis. Figure 5-6 shows where the RBF has placed the four basis functions, at centers C1 to C4 covering four regions.The probability that a customer belongs to a particular region is a function of the distance from the fitting center. Where the basis functions overlap, the probability is calculated as a weighted sum of the functions from the different regions. The weights are calculated to minimize the overall error in classifying all the customers.

In the case of the decision tree classifier we were able to interpret how the classification had been performed by visualizing the decision tree. Interpreting the performance of the RBF classifier can also be done using a visualization technique. In this case we use a visualization similar to that used for the cluster results described in Chapter 4 by replacing the clusters with the RBF regions. As an example the RBF results visualization is shown in Figure 5-7.

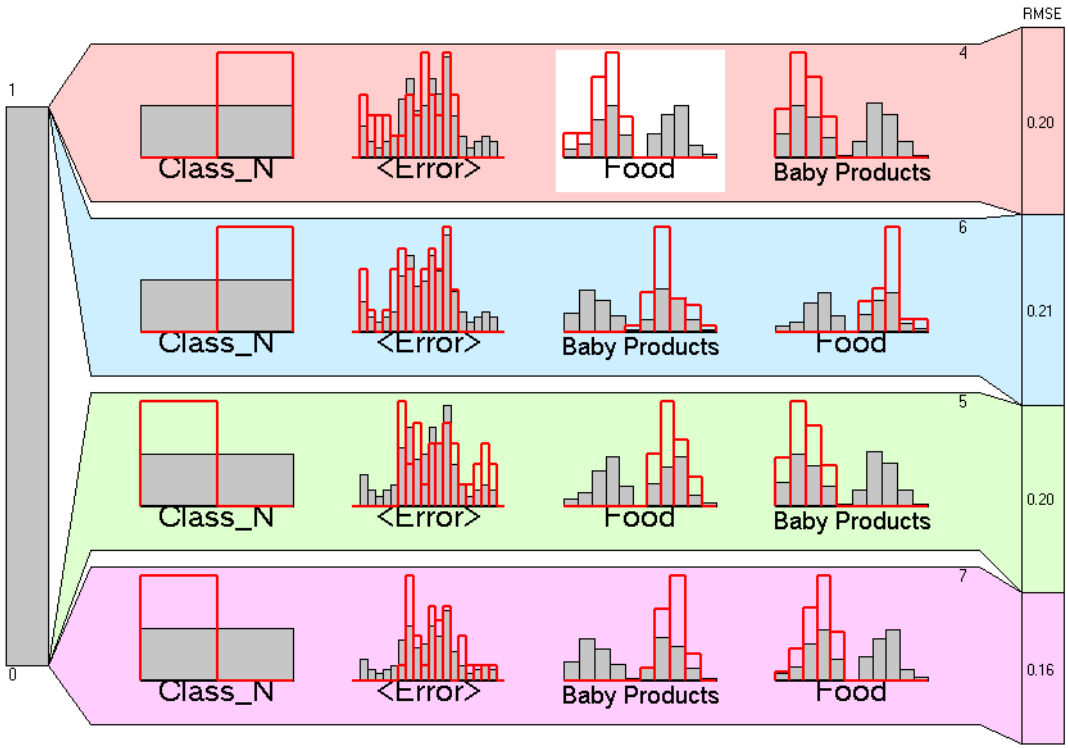


Figure 5-7 RBF results visualization of the regions

Each line now shows a different RBF region and the distribution of the variables of each customer record that are assigned to the region. As with the bivariate and cluster result visualizers, these distributions are compared to the distribution of the variable for all customers. The regions are ordered according to the mean value of the target variable for all customers within the region. This mean value is shown on the left hand side of each line. The first histogram on each line shows the distribution of the target variable for customers in the region. In the case of the top region in Figure 5-7, there are only Family Shoppers and therefore the target variable “Shopper Type_N” only has values of “1”. The mean for this region is therefore 1.0. Similarly the bottom region only has values of “0” and a mean value of 0.0.

As we have already discussed, the predicted value of the target variable within the region depends on the position of the customer relative to both the center of the region to which the customer is assigned and to other regions. The second histogram shows the distribution of the error between this predicted value and the actual value of the variable for these customers. Because we are using the RBF as a binary classifier, the predicted value is actually an estimate of the probability that the customer is a member of the target class.

Note: The value at the right hand side of each line is the Root Mean Square error (RMS) of this error distribution. It should be noted that this is the error around the mean predicted value and not an error on the actual mean value for customers in the region, which is shown on the right hand side of the line. The fact that there is a relatively large RMS error when the RBF is being used as a binary classifier is not unexpected. If there is only one basis function for each distinct group of customers (as in this case), then the probability that a customer belongs to the region is going to vary depending on the distance from the fitting center.

Because the actual value of the target variable can only be “1” or “0”, there will always be an error for each prediction. In this case, where there is little overlap between the basis functions, the error distribution shown for each region, will reflect the distribution of customers around the fitting center (the second histogram on each line). When there is a overlap in the distributions of target categories and non-target categories with similar numbers of customers from the different categories, then the predicted probability can be expected to be of the order of 0.5. In such cases since the target variable is either “1” or “0” we may typically expect RMS errors of around 0.5. We will see this happening with our example customer data set in 5.6.4, “RBF results (TLA model)” on page 126.

The other histograms show the distributions of the customer variables for customers in the region and we can use these to describe the different characteristics of Family Shoppers (top two regions), definite Non-Family Shoppers, which in this case are General Shoppers (bottom two regions).

This type of result visualization therefore helps you to understand the characteristics of the customers in the different regions. If the groups of customers are well separated, then it is relatively easy to understand how a classification decision is made, because the probability that a customer belongs to the region depends on the distance of the customer from the fitting center for that region. However, as we have discussed above, if the different customer groups are close together or overlap, then the probability that the customer is in category “1” or “0” is a weighted sum from the different regions. The region view is then much more difficult to interpret.

An alternative representation is the ‘quantile view shown in Figure 5-8. In this visualization the customers are ordered according to the predicted value which is an estimate of the probability of being in category “1”.

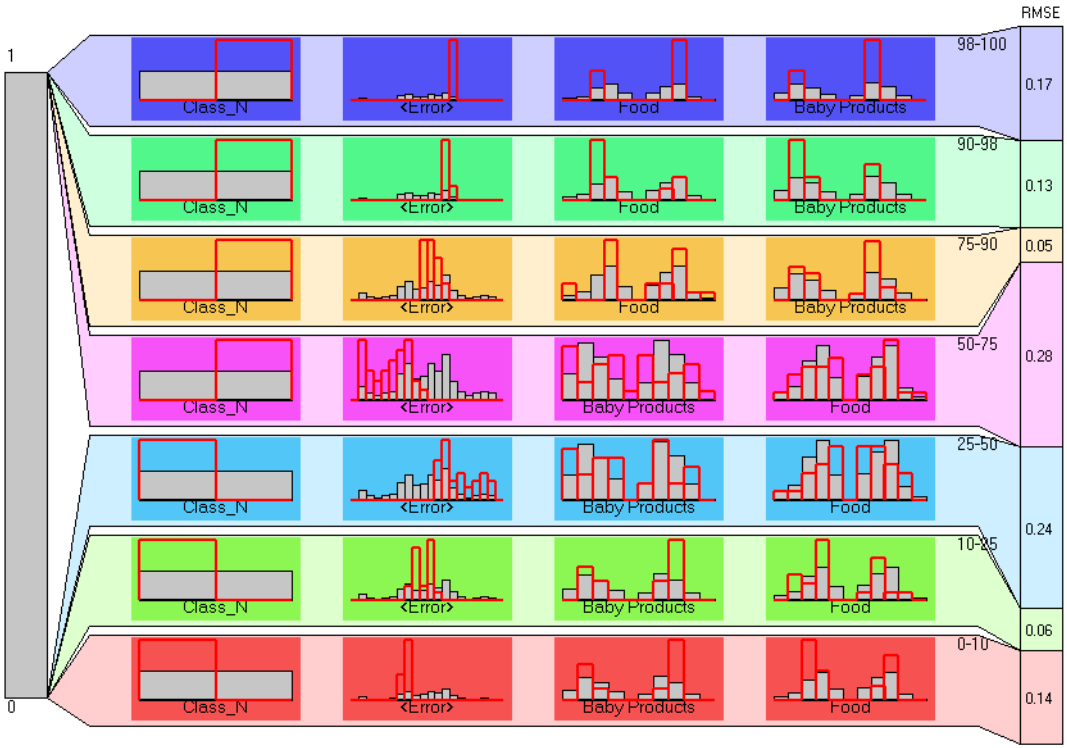


Figure 5-8 Quantile view of the RBF results

The top line now shows the characteristics of the 2% of customers with the highest probability of being in category “1” (98%-100%), the next line shows the customers in the 90% to 98% range and so on. This type of visualization shows the quality of the classification decision and the characteristics of the customers in each of the quantiles, but does not explain how the classification decision was made.

Because the RBF technique is less constrained in the way in which the different regions are constructed, then, in general, it will produce a better result than the decision tree classifier, and particularly where the different customer categories have very similar characteristics. The trade off is again between accuracy and your ability to interpret the reason for the classification decision.

5.5.4 Making decisions using classifier models

A classifier model determines the probability that the customer belongs to each of the possible categories into which they can be placed. To make a decision about which category to assign the customer to is then usually a question of selecting the category with the highest probability (usually termed the *winning class*) and declaring the customer to belong to that category. However, there are situations where the classification probability is not sufficiently high to make the decision and in these cases you have to be content with an “unknown” classification. Where you have multiple categories then you know that the minimum probability for the winning class must exceed the value of one divided by the number of categories. Similarly, where there is a risk of incorrectly classifying, and particularly where you have used error weighting to account for this risk, it is important that you define the minimum threshold to be one minus the risk probability. This idea is illustrated in Figure 5-9.

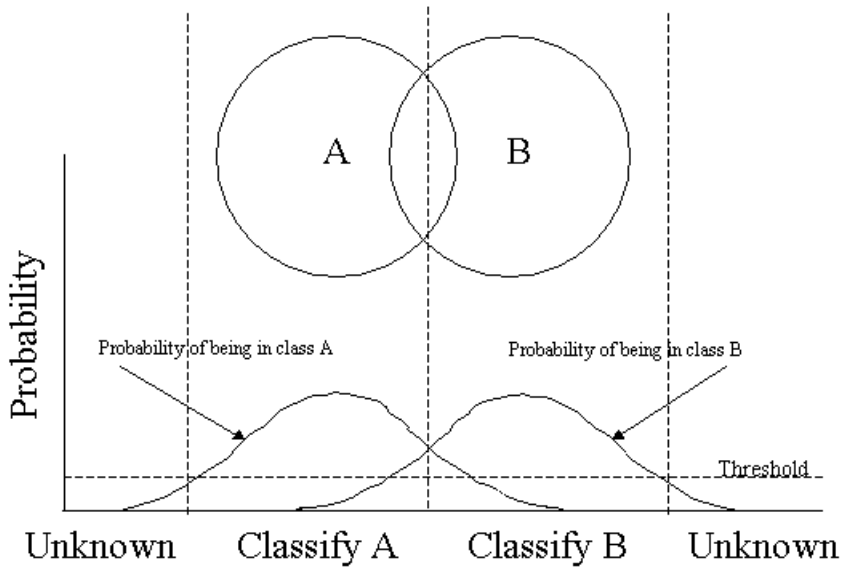


Figure 5-9 Classification decision and the unknown class

When using decision tree classifiers it is important to recognize that the minimum threshold will always be exceeded, but that if a risk threshold is specified then it is possible to get a classification confidence below the risk threshold. You should also note that classification rules are open ended, and therefore when performing classification it is possible to obtain a high classification probability for data records whose variables (that are being used to perform the classification) are outside of the range of the training group. You should therefore ensure that the data being by the classifier is within range.

This is not the case for the RBF classifier, where for data records far from the fitting centers, the classification probability tends towards zero, and in these cases it is possible to get a probability below the minimum threshold. Because it is difficult with the RBF classifier to determine which features are being used to perform the classification, it is not always possible to check that the data you are using is within range. Using the minimum threshold to some extent removes the requirement to make these checks.

5.6 Interpreting the results

Using the CLA and TLA data models, the different classification techniques can be used to construct classification models for our example data set. This section describes the performance of the decision tree and RBF classification models and how the results can be interpreted. This is the *sixth stage in our generic data mining method.*

5.6.1 Decision tree classifier (using CLA data model)

You can construct a classification model based on the CLA data model by using: the NRS of customers' measurement aggregated at the products group level as the input variable to the model, and the business segment name as the target variable to the model. The resulting decision tree using the example data set is shown in Figure 5-10.

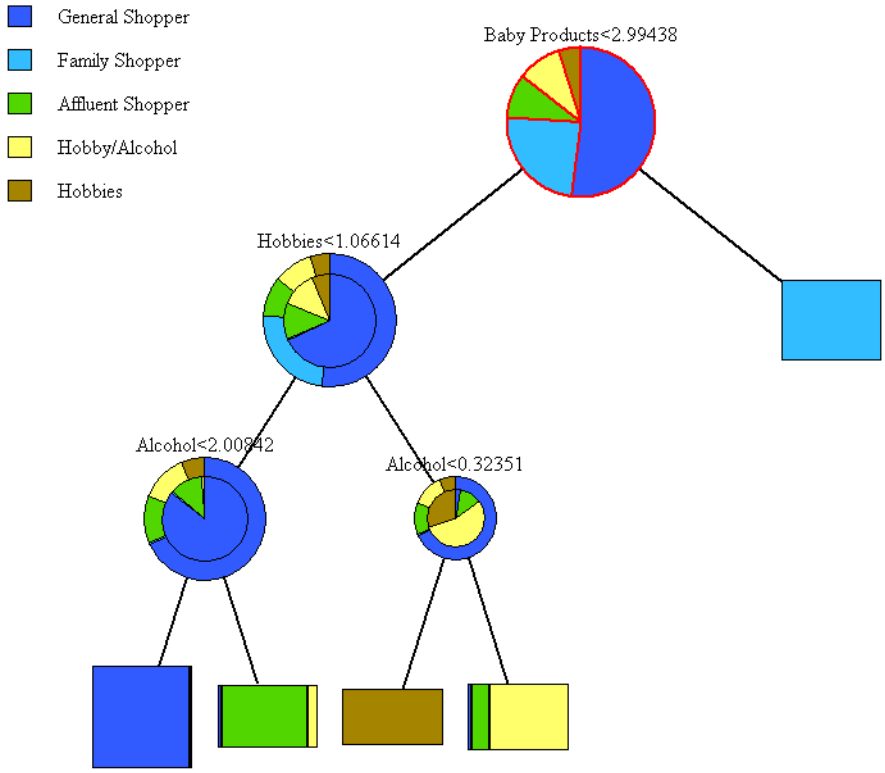


Figure 5-10 Decision tree for the CLA data model

The tree shows that by using the aggregated NRS of the CLA data model, customers can be classified into the appropriate business segments with almost no errors for the General, Family and Hobby Shopper categories and just a few errors for the Hobby/Alcohol and Affluent Shopper categories. A statistical summary of the actual errors that are generated for the training data set and then the test data set is given in Figure 5-11 and Figure 5-12.

```
Train NRS
---------
Number of classes = 5
Errors            = 12 (4.8%)

Confusion matrix for pruned tree

Predicted Class --> | General Shoppe | Family Shopper | Affluent Shopp | Hobby/Alcohol  | Hobbies        |
                    ------------------------------------------------------------------------------------
 General Shoppe     |            128 |              0 |              1 |              1 |              0 |    total = 130
 Family Shopper     |              1 |             59 |              0 |              0 |              0 |    total = 60
 Affluent Shopp     |              2 |              0 |             17 |              5 |              0 |    total = 24
 Hobby/Alcohol      |              0 |              0 |              2 |             22 |              0 |    total = 24
 Hobbies            |              0 |              0 |              0 |              0 |             12 |    total = 12
                    ------------------------------------------------------------------------------------
                               131 |             59 |             20 |             28 |             12 |    total = 250
```

Figure 5-11 Confusion matrix for CLA data model training set

```
Test NRS
--------
Number of classes = 5
Errors            = 16 (6.426%)

Confusion matrix for pruned tree

Predicted Class --> | General Shoppe | Family Shopper | Affluent Shopp | Hobby/Alcohol  | Hobbies        |
                    ------------------------------------------------------------------------------------
 General Shoppe     |            116 |              0 |              0 |              1 |              3 |    total = 120
 Family Shopper     |              2 |             62 |              0 |              0 |              1 |    total = 65
 Affluent Shopp     |              4 |              0 |             18 |              4 |              0 |    total = 26
 Hobby/Alcohol      |              0 |              0 |              1 |             25 |              0 |    total = 26
 Hobbies            |              0 |              0 |              0 |              0 |             12 |    total = 12
                    ------------------------------------------------------------------------------------
                               122 |             62 |             19 |             30 |             16 |    total = 249
```

Figure 5-12 Confusion matrix for CLA data model test set

The two displays shown above are termed “confusion matrices”. They each show the numbers of errors that result from the classification for each of the target classes. In the case of the training set the overall error is 4.8%, increasing only slightly to 6.4% for the test set. This excellent result should not be surprising, because we know from the previous chapter that the original assignment of customers to clusters was based on an assessment of the NRS.

Using the rule generation capabilities of the decision tree classifier, we can examine how particular classification decisions are made. An example is shown in Figure 5-13.

Figure 5-13 Classification rule example

The business rules used to perform the initial classification were more qualitative than the rule shown in Figure 5-13, for example the corresponding business rule was:

- General shoppers: Characterized by spending mainly on Food and Household goods with a relatively small expenditure on Alcohol and Baby Products and Hobbies.

The decision tree rule confirms this definition but has now quantified what the term “relatively small” actually means for each of the three product groups. The technique has therefore discovered a quantitative method for performing the classification and provides a way of accurately classifying new “loyalty” card customers to the business segments.

5.6.2 Decision tree classifier (using TLA model)

While the above result confirms that the classification technique works well for the aggregated customer data, these customers had to be loyalty card customers and may only represent a small proportion of the total customer base. The main objective in this example is to address the question of whether your customers can be classified into the same business segments using purchase record data from single transactions.

As we discussed in the previous chapter, if a customer purchases many articles during a single transaction, then the aggregated NRS and the single transaction NRS are likely to be similar and a good classification performance would be expected. When only a few items are purchased, however, it is less obvious what the result will be.

To address this question you can use the NRS from the TLA model aggregated at the product subgroup level for single transactions and use this data for the training group of customers to build a decision tree model. The resulting decision tree using our example data set is shown in Figure 5-14.

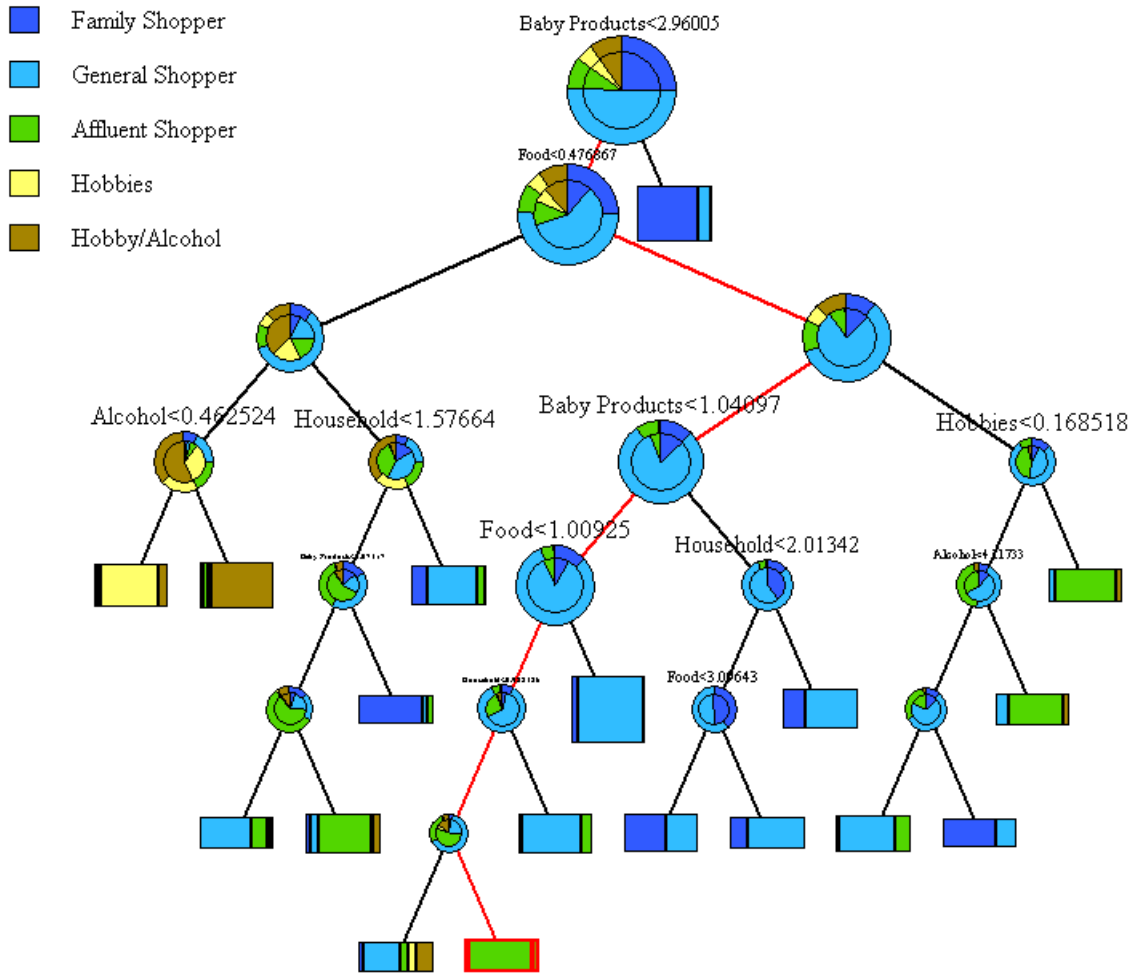


Figure 5-14 Decision tree for the TLA data model

Although the tree is now much more complex, reflecting the greater variability of the NRS from single transactions, in judging the quality of the tree, you should note that it is well balanced, not too deep and makes what appears to be sensible splits. The tree shows that for some customer categories, and for a subset of customers in these categories, the classification performance is very good (for

example, some of the Family Shopper and Affluent Shopper customers). Each of these leaf nodes can be described by a rule that identifies the characteristics of the customers at the node. Therefore, customers with these characteristics will be classified with a high probability. At some leaf nodes there are still a mixture customer from different categories. For customers with these characteristics, it will be difficult to classify and this is reflected in a lower probability. Figure 5-15 and Figure 5-16 show the corresponding confusion matrices for the test and training sets.

```
Classify Shoppers
-----------------
Number of classes = 5
Errors            = 473 (18.92%)


Confusion matrix for pruned tree

Predicted Class --> | Family Shopper | General Shoppe | Affluent Shopp | Hobbies        | Hobby/Alcohol  |
                    ---------------------------------------------------------------------------------------
 Family Shopper     |            464 |            146 |              6 |              5 |              3 |    total = 624
 General Shoppe     |            122 |           1115 |             19 |              5 |              6 |    total = 1267
 Affluent Shopp     |              5 |             77 |            132 |              5 |             16 |    total = 235
 Hobbies            |              0 |              3 |              1 |            120 |              7 |    total = 131
 Hobby/Alcohol      |              0 |              8 |             17 |             22 |            196 |    total = 243
                    ---------------------------------------------------------------------------------------
                               591 |           1349 |            175 |            157 |            228 |    total = 2500
```

Figure 5-15 Confusion matrix for the TLA data model training set

```
Classify Shoppers Apply
-----------------------
Number of classes = 5
Errors            = 546 (21.84%)


Confusion matrix for pruned tree

Predicted Class --> | Family Shopper | General Shoppe | Affluent Shopp | Hobbies        | Hobby/Alcohol  |
                    ---------------------------------------------------------------------------------------
 Family Shopper     |            452 |            157 |              2 |              3 |              1 |    total = 615
 General Shoppe     |            149 |           1048 |             31 |              7 |              2 |    total = 1237
 Affluent Shopp     |             21 |             88 |            130 |              1 |             19 |    total = 259
 Hobbies            |              0 |              3 |              1 |            112 |             10 |    total = 126
 Hobby/Alcohol      |              0 |              8 |             17 |             26 |            212 |    total = 263
                    ---------------------------------------------------------------------------------------
                               622 |           1304 |            181 |            149 |            244 |    total = 2500
```

Figure 5-16 Confusion matrix for the TLA data model test set

The overall error on the training set data is now 18.9%, and on the test set data it is 21%; although as the tree itself shows, there are some “leaves” where the classification performance is much better than at other leaves. As expected, the performance is not as good as for the CLA data, but as we will see this is still an acceptable level of performance for some types of applications.

The next question is, How do you assess how good the classification performance is for an individual business segment? One important technique that can be used to do this is called the gains chart and this is described in the next section.

5.6.3 Measuring classification performance (gains charts)

The concept behind the gains chart is one of ordering, or ranking customers in either the test or training data sets based on the confidence that they belong to the target class. For example, suppose that we wanted to rank our customers in this way for the General Shopper class. One possible ranking of the our customers is to simply guess the order and arrange our customers in a list according to our guess. To measure how well we performed, we could then start with the customer at the top of the list and look at the actual class to which they belong. If we guessed correctly, we give ourselves a score of “1”, and “0” otherwise. As we move down the ranking, we sum the score until we come to the end of the list. If we were to plot the score against the ranking of customers, we would get the curve shown as (a) in Figure 5-17.

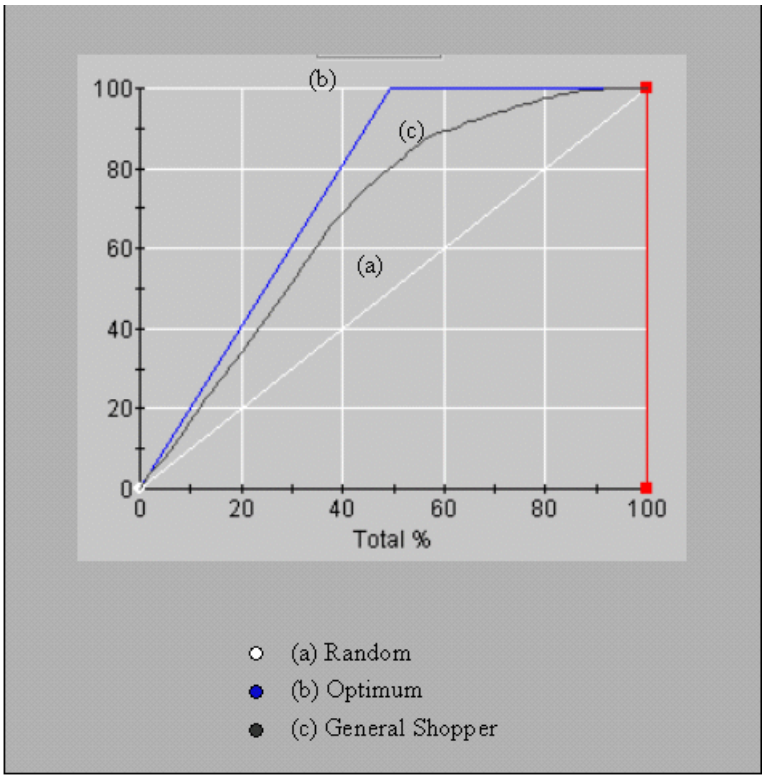


Figure 5-17 Gains chart for the General Shopper classification

Curve (a) shows that on average we would correctly identify our General Shoppers in direct proportion to the number of customers in the list. To correctly identify all our General Shoppers, we would have to classify all our customers as General Shopper and accept that in this case we would end up with 50% of customers incorrectly classified.

An alternative ranking could be obtained by using the actual customer segments and then ranking them with the General Shoppers first and repeating the scoring exercise as we moved down the list. In this case we would get the curve (b) shown in Figure 5-17. This is the optimum result that we could obtain, because it is based on a perfect knowledge of the customers.

If we used our classification model to rank the customers, where the confidence of being a General Shopper is determined from the distribution of Shopper Types in the leaf nodes of the tree, then we would get a curve similar to (c) in Figure 5-17. If we had leaf nodes that only had General Shoppers, then our confidence would be 100%, and these customers would be ranked highest in our list so that initially curve (c) would follow the optimum curve (b). As we move to leaf nodes that contain a mixture of General Shopper and other shopper types, our confidence decreases and we begin to introduce errors in our classification; and so we fall below the optimum curve as shown. If our classifier had been perfect, then curve (c) would have been identical to curve (b). If our classifier had produced a random decision, then it would have matched curve (a), and if it had been incorrectly biased it could have been below curve (b). The measure of how much better our classifier is from a random guess is termed the “Lift” or the “Gain” and therefore the name Gains or Lift chart for diagrams like Figure 5-17.

Using the results from the TLA decision tree classifier, each of the five Shopper Types results in the series of gains charts shown in Figure 5-18.

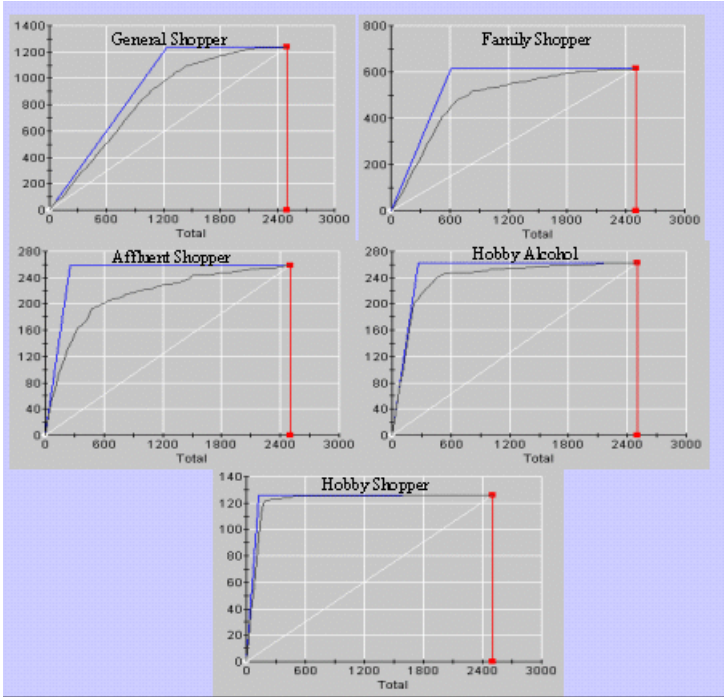


Figure 5-18 Gains charts for each of the five Shopper Types using the TLA model

Because we have used the TLA data model to build the classifier, in this case the gains chart is obtained by ranking the transactions rather than the customers. The results show that the classification can be performed with a high degree of confidence for all categories, and that in the case of the Hobby Shopper and Hobby/Alcohol Shopper, the classifier model performance if near the optimal.

The ways in which we use gains charts to compare the performance of different classification models is described in 6.5.5, “Generating scores including association rules” on page 158 and the way in which they can be used to target customers is described in 6.7, “Deploying the mining results” on page 167.

5.6.4 RBF results (TLA model)

Using the RBF technique, you can also construct classification models using both the CLA and the TLA data model as for the decision tree classifier, but now includes the additional binary variable for each of the Shopper Types. In this case we have used the TLA model and the example data set to produce an RBF classifier for each of the five shopper types. The resulting models can be visualized as we described in 5.5.3, “Radial Basis Function (RBF)” on page 111, and Figure 5-19 shows the result for the General Shopper Type classification.

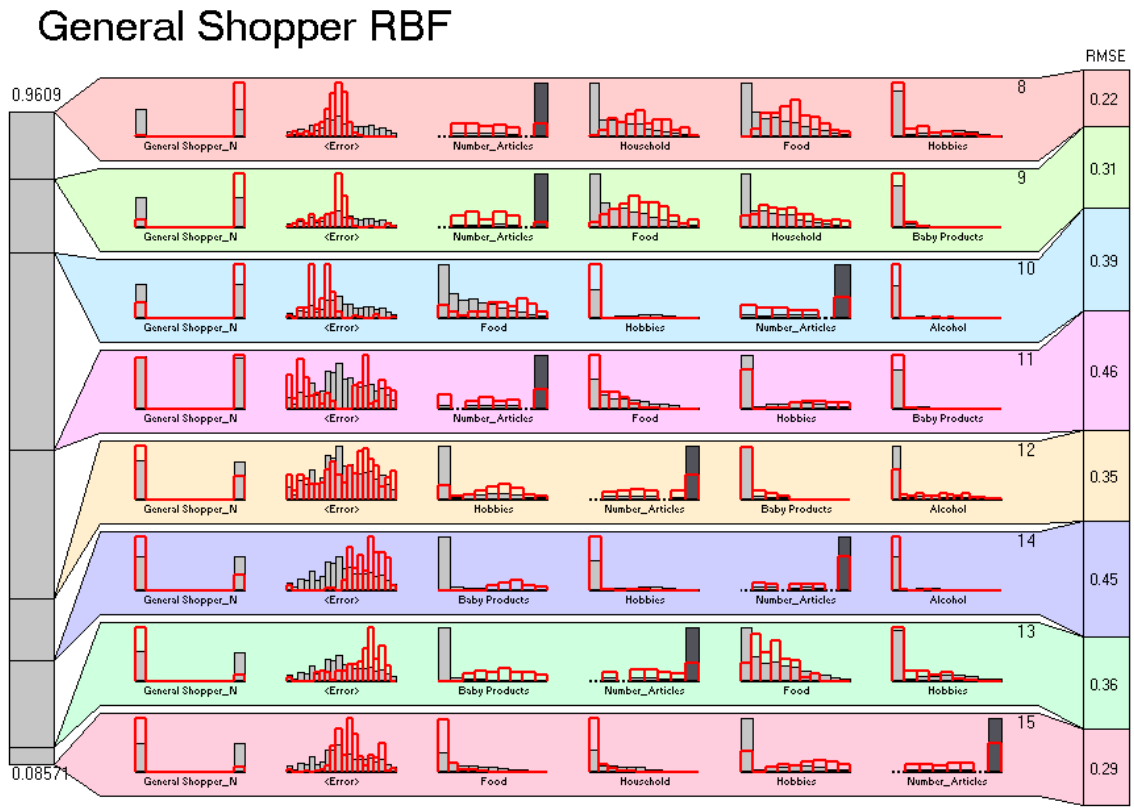


Figure 5-19 RBF visualization for the General Shopper classifier

Figure 5-19 shows eight different RBF regions. The number of regions is determined either by specifying the minimum number of customer transactions that should be used to define a region or by specifying the number of regions required. In this case, after some investigation of the stability of the prediction for different numbers of regions, we chose eight regions. The results show that using the NRS from single transactions, the top three regions are still predominantly General Shoppers with the bottom three regions being predominantly Non-General Shopper. The middle two regions are indeterminate. Note that in this case where there is the most uncertainty, the RMS error is 0.46, which is what we expected to find if the predicted value was around 0.5.

In this case, the regions themselves give a good indication of the classification, and this is confirmed in the “Quantile” view of the results shown in Figure 5-20.

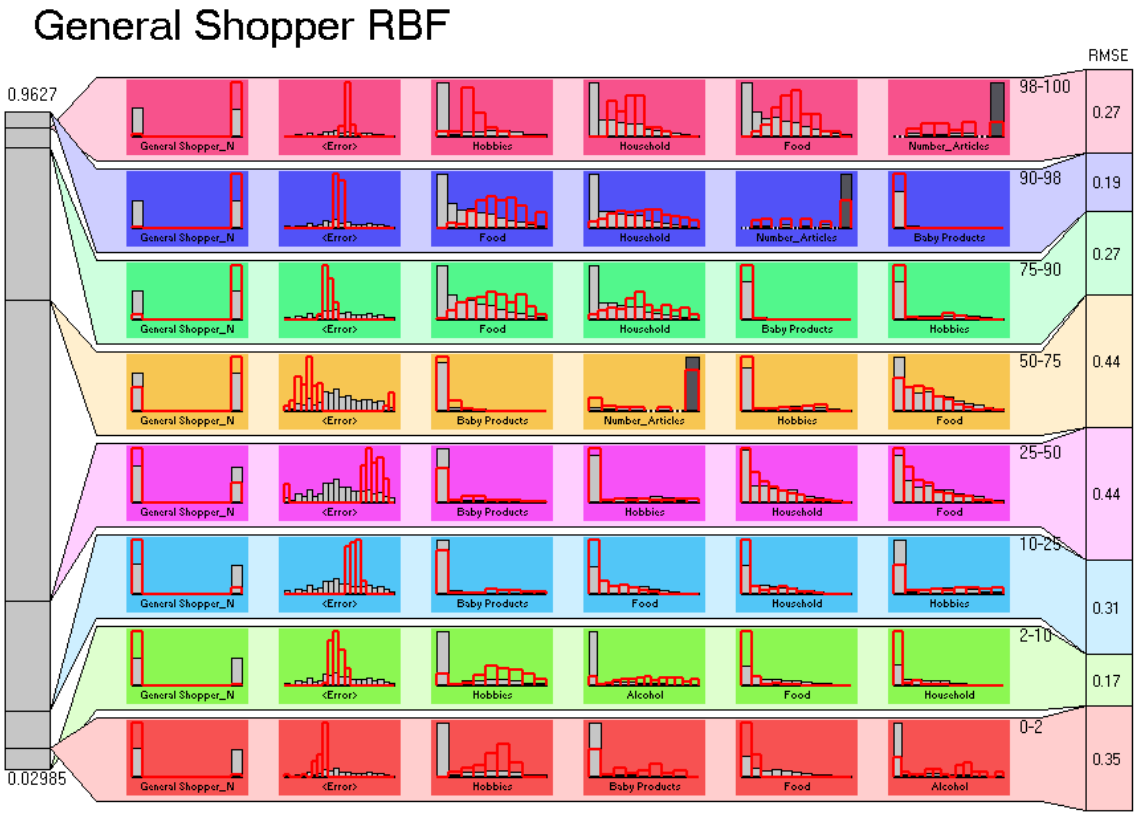


Figure 5-20 Quantile view of RBF results for General Shopper type

As we explained in 5.5.3, “Radial Basis Function (RBF)” on page 111, the quantiles are identified by the percentile ranges on the right hand side of each line. The customers in each quantile are determined from the predicted value of the target variable, and the mean value of the actual value of the target variable is shown on the right hand side. The large RMS error of the middle quantiles is around 0.5, as we expect. Again the predicted probability in these quantiles will be around 0.5, and there is a mixture of shopper types in these quantiles (seen from the first histogam in each row). Similarly, the average RMS error of 0.25 for the upper three quantiles (75% - 100%) is due almost entirely to a predicted probability of around 0.75, because the majority of customers in these quantiles are General Shoppers. Similar reasoning leads to the interpretation of the lower three quantiles (0% - 25%), where the predicted probabilities are around 0.23 with an average RMS error of 0.27 with these regions comprising mainly Non-General Shoppers. We therefore expect that the prediction performance of the model will be very good, and this is confirmed by using the gains charts.

Performing the RBF classification for each of the shopper types and using the gains charts for each, enables us to compare the RBF results. These are shown in Figure 5-21.

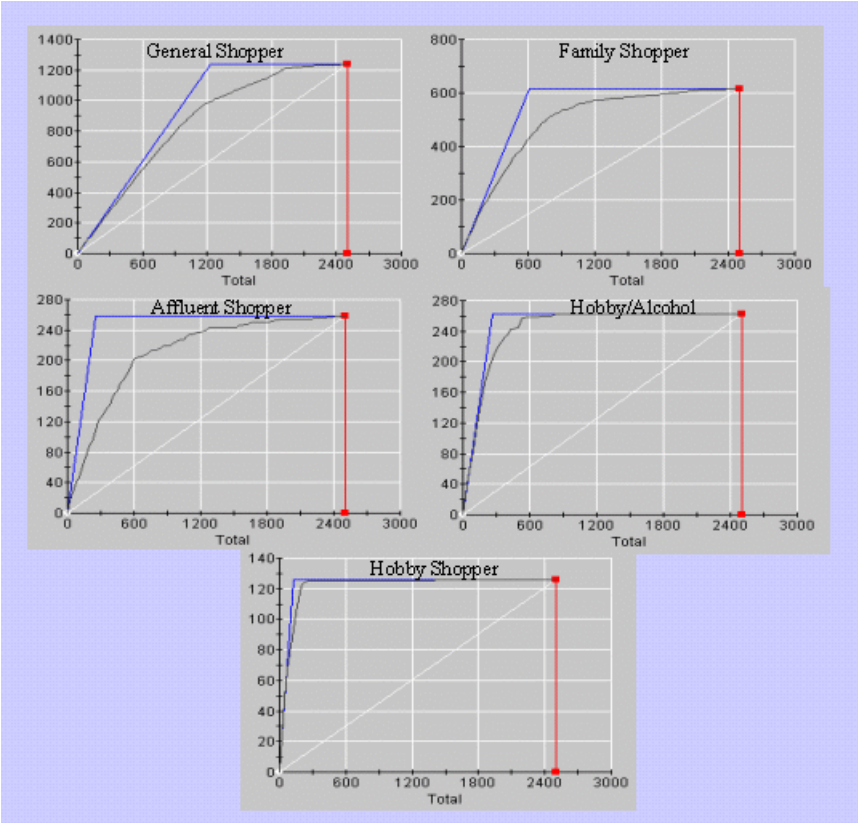


Figure 5-21 Gains charts for the RBF classifier using the TLA model

Again the results show a very high level of classification performance, particularly for the Hobby and Hobby/Alcohol Shoppers.

5.6.5 Comparison of the decision tree and RBF results

The *decision tree* and *RBF* results can easily be compared using the gains charts and series of comparisons for the different shopper types is shown in Figure 5-22.

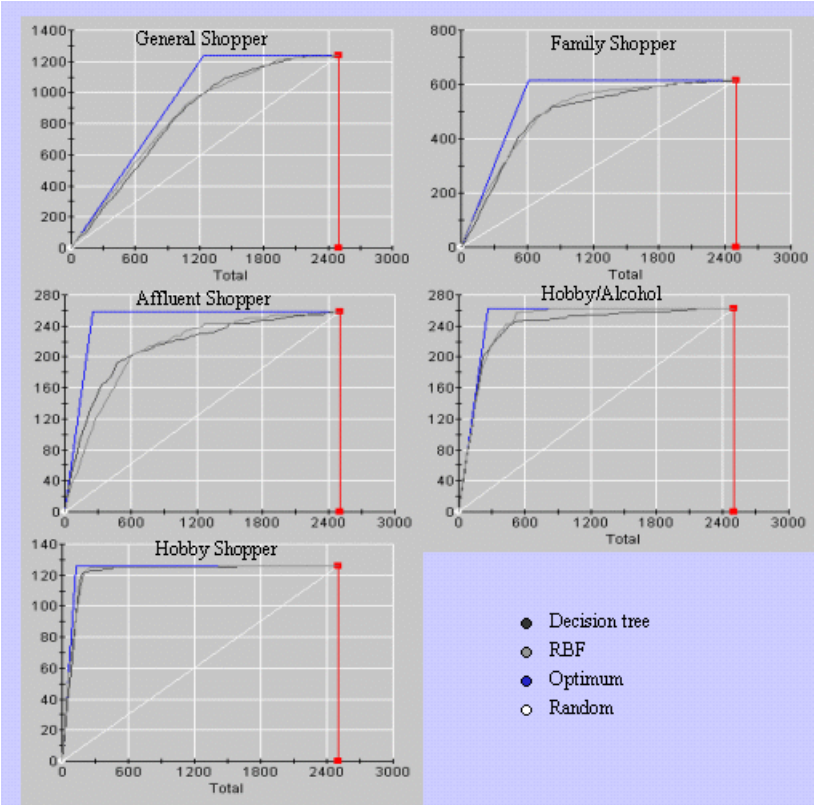


Figure 5-22 Comparison of the decision tree and RBF classifiers for the TLA data model

In the case of the General Shopper class, the RBF and decision tree results are very similar, whereas for the Affluent Shopper class the decision tree initially outperforms the RBF, but beyond the 800 transaction point the RBF result is better. For the other classes, the RBF consistently out performs the decision tree. This result is as expected, because the RBF classifier is able to divide up the variable space without being restricted to using a single variable at each split, and in general this additional flexibility leads to a better classification performance.

There are a number of ways in which the two classification results can be combined to produce an overall result. If you are particularly bullish you can classify each customer by taking the classifier result with the highest confidence. If you are pessimistic you can do the converse. In general, and particularly if you are cautious, then you would perform a simple average of the predicted probabilities for the two results.

5.7 Deploying the mining results

There are many potential ways of deploying the classification models into your retail business. This final and *seventh stage in our generic mining method* is once again a crucial step that needs to be carefully planned and executed if the full value of performing the data mining is to be released. In this section we will look at two possible ways in which the classification results obtained from our example data can be used. These are not the only possibilities, but they do indicate the different approaches that can be used and the sorts of things that you need to think about. In the first example we look at how classification results can be used as part of a directed marketing campaign. In the second example we consider how classification can be used at the point of sale, or in store kiosks or other customer touch points.

5.7.1 Direct mail and targeted marketing campaigns

A direct mail or targeted marketing campaign assumes that you are able to contact your customers directly in some way. If you want to focus your campaign around the type of market segments that we have been discussing, then you must have some way of linking the customer to the transaction data that was used to define the segments. If this linkage allows you to aggregate transaction records, then you can use the CLA data model to construct your classifier and take advantage of the improved performance that this affords.

What you want to be able to do is to direct your campaign to the customer segment that is most appropriate and to maximize the return on investment of the campaign. For example, suppose that you wanted to target the Family Shopper customers with some new type of baby product. You may already know from some type of focus group activity, that the best type of campaign is to directly mail free samples, because this has a high take-up rate (say 10%), but only for Family Shoppers. In this case you clearly want to ensure that you target the Family Shopper segment, because this type of campaign has relatively high costs associated with it, and you want to avoid mailing the offer to Non-Family Shoppers. To assess how successful your campaign is likely to be, you can use the gains charts that we described earlier. If we rank each of our customers in terms of the probability that they will be a family shopper, and then assume that we will selectively mail the highest ranking customers, we can calculate the cost of the campaign and the predicted Return On Investment (ROI).

Ideally, you would like to mail all Family Shoppers and no other shopper types. This would then maximize the ROI. If you mail all your customers, you will clearly reach all of the Family Shoppers with the free sample, but at the cost of having to mail everyone. The question now is how many customers do you mail to, to maximize the ROI? To answer the question we need to know the following:

- The cost of mailing a free sample to each customer
- The predicted revenue (or more likely a range of predicted revenue) that a positive response to our offer will generate
- The probability of response to the campaign if we correctly mail to a Family Shopper rather than any other type of shopper
- The rank of each customer as a member of the Family Shopper category (from our gains chart table)

Using this information we can calculate the cost and predicted revenue from mailing to customers based on mailing our customers in the order of the ranking that we have given them, and therefore determine how many customers we need to mail to maximize our profit. Figure 5-23 shows the gains chart that we have used to perform this calculation, and Figure 5-24 shows the corresponding profit that would be generated assuming a mailing cost of $1 and for different levels of revenue that would result from each positive response.

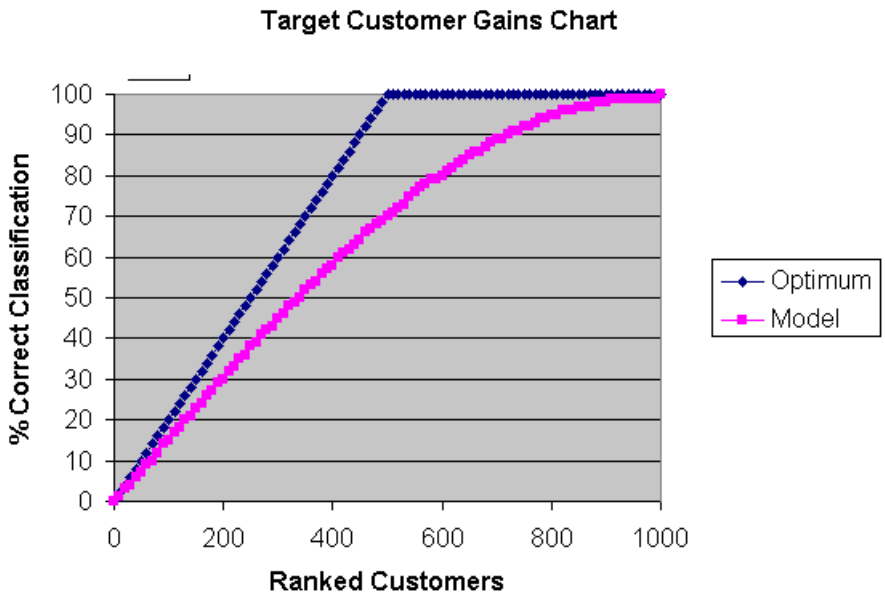


Figure 5-23 Target Customer Gains Chart

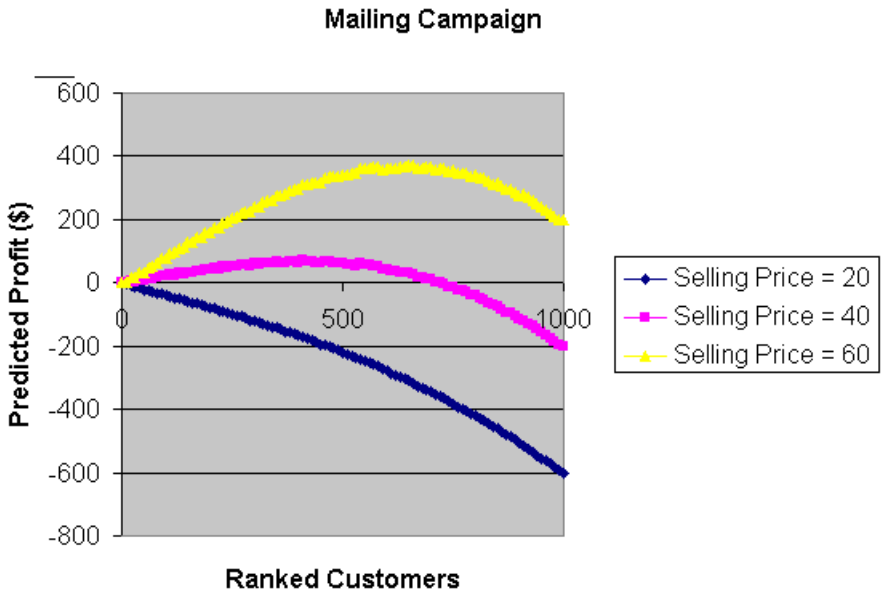


Figure 5-24 Predicted Profit from campaign

The results show that if the predicted revenue from each positive response was $20, then we would not make any profit irrespective of how many customers we mailed. At a revenue value of $40, we could maximize our profit if we mailed only 40% of our top ranked customers, but would make a loss if we mailed more than 75% of our ranked customers. At a revenue value of $60, we would always make a profit from the campaign, but this could be maximized if we only mail 60% of the top ranked customers.

These types of calculations can be performed using simple spreadsheet models, or by scoring the customers and performing the ranking necessary calculations in the database, or by making the scores or the classification model available to any CRM tools that perform this type of function. What is important to recognize is that it is the quality ranking rather than the absolute value of the probability of the classification that drives this type of process.

Using classification in this way provides a means of quantifying the predicted return on investment from a marketing campaign and greatly enhances your decision making process about which campaigns are appropriate in any given situation.

5.7.2 Point Of Sale and kiosk offers

In the previous chapter we looked at scoring customers using the results of clustering and using this information to promote products at the point of sale or in store kiosks through other customer touch points. Rather than use clustering models to do this, you can now score your customers using the classification models (the classifiers are also supported by IM Scoring to do this). The main advantage of using classifier models is that, as we have shown, it is possible to construct reliable models that can classify your customers using data collected at a single transaction. This enables you to perform the scoring directly at the point of sale without the requirement to have access to your customer database. This concept is shown in Figure 5-25.

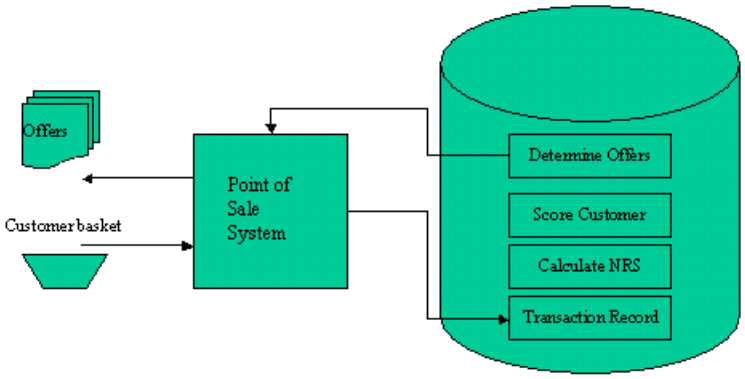


Figure 5-25 Point of sale system using classification scoring

If you want to extend this capability to a kiosk or other customer touch point then there are a number of options. One possibility is to use the transaction receipt itself. The transaction receipt can be read by a device in the kiosk and using the information, the customer can be immediately categorized and appropriate offers made. Alternatively, the customer could be issued with some form of ticket that has the transaction data encoded on it (either as a bar code or on a magnetic strip) and this could be used in a similar manner. This concept is shown in Figure 5-26.

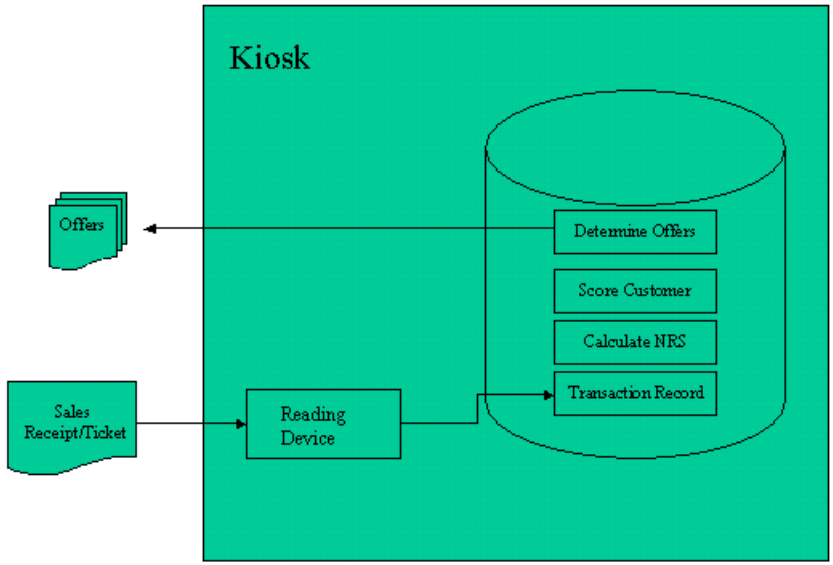


Figure 5-26 Scoring customers at a kiosk

There are a number of alternatives to this type of system including recording the information on a customer loyalty card but this will depend on which customers you want to target with the offers. The issue now becomes that of deciding which products or services to offer to customers in a particular segment and how to make these offers relevant and exciting. This is the subject of the next chapter, Chapter 6, “How can I decide which products to recommend to my customers?” on page 137.

6

How can I decide which products to recommend to my customers?

With a large range of diverse products to offer, and with customers purchasing from a wide range of product groups, a major challenge for any retailer is to decide how you select the most appropriate products to recommend to your customers?

In this chapter we describe how data mining can be used to identify *cross-sell* and *up-sell* opportunities within your retail business. This is achieved by matching the expected appeal of a product, not previously purchased by a customer, to the spending by the customer on related products. Associations mining is used to determine product appeal and this is combined with clustering analysis to identify specific product items to recommend. The technique can be used to automate personalized product recommendation systems, or to determine product placement within retail outlets.

6.1 The business issue

Every retail organization wants to encourage its customers to widen their relationship, through the purchase of new products and services, or to deepen their relationship, through the purchase of more profitable products and services. A campaign that encourages customers to purchase new products and services is traditionally termed *cross-selling*, while a campaign that encourages the purchase of more profitable products and services is termed *up-selling*. The challenge is to determine which type of campaign is appropriate to a specific customer and which products or services are the "best" to recommend.

In this context by "best" we mean that the recommended product is not only a good match to the customer, but also that the recommendation will in some way be relevant and unusual or interesting to the customer. By being relevant the customer identifies the offer as being personalized, and by being unusual or interesting, hopefully this will excite the customer to purchase the recommended product. As an example of this, we might identify a customer for whom a good product match may be fresh vegetables, although this customer does not normally purchase products from this category. Offering artichokes to this customer, rather than say turnips, would be a more interesting and unusual choice.

Where there are a limited number of products and services, the identification of cross-sell or up-sell opportunities is often intuitive and similar to our market segmentation example (in Chapter 4), and we could develop some simple rules for making appropriate offerings. However, the challenge for the retailer is that with large numbers of products and services that could be recommended (for example, a supermarket may have 10's of thousands of individual products), which are the "best" ones to promote to particular customers?

The recommendation problem is compounded by the dynamic nature of a typical retail business. Products are continuously being replaced with new products and there is an increasing diversity of routes through which you interact with your customer. With the advent of e-commerce, these routes are, themselves, becoming increasingly dynamic, and customers using these services want new and interesting ideas each time they use them. To keep pace with this change, product recommendation systems must become equally dynamic. This requires a capability to automatically react and modify the recommendations that are made in response to the changing environment. How can data mining provide a solution to this problem?

6.1.1 What is required?

Applying the *first stage in our generic data mining method* requires us to translate the business issue into a question, or set of questions, that can be addressed by data mining. To address the specific question of identifying products or services might be most appealing to a particular customer there there are a number of approaches that could be taken.

As discussed in Chapter 4, one approach is to group customers who exhibit similar buying behavior, and then develop campaigns that can be targeted at the group rather than the individual. In a similar way, we can identify customers and the products they purchase, and then offer cross-sell, or up-sell suggestions from products purchased by customers within the same segment.

As a simple example of this type of cross-sell suggestion, suppose we have a segment that we identify as Breakfast Shoppers. If one customer in this segment purchases bread, butter and preserves (Customer A) and another purchases bread, butter and honey (Customer B), then offering honey to A and preserves to customer B would seem to be a sensible cross selling suggestion. The problem with this approach is that if offering preserves to Customer A causes this customer to change their eating habits and simply stop buying honey and start buying preserves, while Customer B does the opposite, we would have achieved precisely nothing. This type of product substitution is often termed *cannibalism*. What we need to do in this case is to offer a product to both that is not a substitute for one product or the other, but truly broadens the customer relationship, for example an electric toaster.

A simple example of up-sell would be where Customer A purchased a more profitable brand of butter than Customer B. Promoting the more profitable brand of butter to Customer A would seem to be a good up-selling opportunity. There are a number of point of sale systems that essentially use this approach; for example, if a customer buys a particular brand of a product then promote the more profitable, usually a more expensive brand to this customer with a discount coupon to encourage the initial purchase. You don't need data mining to do this type of up-selling, but it does not produce particularly interesting alternatives, neither does it recognize that other types of purchases may be indicative of the customers propensity to purchase more expensive brands.

In data mining terms this type of approach to product recommendation is called *content based filtering*. Where there are many possible products to choose between, then even this simple approach to cross-selling and up-selling becomes a challenge. The solution that can be used to solve the challenge of the many product problems is perhaps most familiar in the use of the technique for document retrieval systems, such as the Web search engines. In this case a list of key words specified in the search is matched with the occurrence of the same key words in the document database and the documents with the highest degree

of match are returned. A similar approach can be used to product recommendation by replacing the key words with the products and the documents by groups or segments of similar customers. The problem with using this type of content based filtering in a retail environment is that, although customers are grouped because they have a preference for specific product groups, at the same time they still purchase products from most of the other categories. This is rather like most of our documents having all of the same key words, but some documents use some key words more than others. If we simply select products on the basis of the most popular products in the customer segment, we may miss some important combinations of products that are purchased by other customer groups. If we want to maximize our cross-selling potential, there is much that can be learned from the purchasing behavior of customers in other segments and we need to make use of this information.

An alternative way of identifying products is called *collaborative filtering*. The concept behind collaborative filtering is to use information about the purchasing behavior of all of our customers and to match products to customers based on the expected appeal of the product. The expected appeal is derived by looking at how products are purchased in combination by all customers and then by recommending new products to customers who purchase subsets of the combinations, but not the recommended products. As an example, suppose we have a product such as a particular type of wine, and we identify that there is a high likelihood that if customers purchase sports products, they will also purchase wine. If we have a customer who has a high spend in sports products, but does not normally purchase wine, then by association there is an indication that wine products should be recommended to this customer. If the customer also purchases cheese and there is a similar likelihood that cheese and wine are associated, then there is an even stronger case for recommending wine to this customer. What we seek to achieve by collaborative filtering is to score for each product in terms of its expected appeal and then offer the customers a list of recommendations from the highest scoring products.

6.1.2 Outline of the solution

To help you understand how a collaborative product recommendation system works and the role data mining plays we need to consider what the main components of such a system may be. It turns out that we need four main components:

- A way of characterizing our customers in terms of the products that they purchase
- A way of determining the significant combinations of products that are purchased across all of our customers

- A way of combining the two types of measurements to produce a score that matches customers to products
- A way of ensuring that the recommendations are appropriate and are in line with our business objectives

The approach we use builds on a data mining technique described in the research paper, *Personalization of Product Recommendations in Mass Retail Markets*, which combines the advantages of both the content and collaborative approach to personalization. (See “Other resources” on page 183 for the bibliographic information for this book.)

In this approach, customers are characterized by their spend in different product subgroups. In Chapter 4, we described the concept of Normalized Relative Spend (NRS) as a measure of customer expenditure and used the NRS, aggregated at the product group level of the product hierarchy to develop data driven customer segmentation. For the purposes of the product recommender, we also use the NRS, but this time aggregating at the product subgroup level. The primary reason is that in most retail organizations individual products are often replaced or substituted on a regular basis, and it is often difficult to construct a stable picture of what is happening at the product level. Although this can be true at the subgroup level (or their equivalent in your organization), this tends to remain stable over much longer periods of time and this is directly related to the stability of the recommendations made.

Matching customers to products will require us to create what we term “customer records” and “product records”. Customer records comprise the NRS in each product subgroup and have an entry for each subgroup. As we will show, the product record is derived from the data mining, so that, like the customer records, it has an entry for each product subgroup. We will describe in some detail how this is done later in the chapter. The important thing is that because the two types of records have the same structure, they can be compared and a score produced for every product and every customer.

To produce product records, we need to discover the significant combinations of products that are purchased by all of our customers. To do this, we will use a technique known as associations rules, and to match with our customer records, the product associations will be determined at the product subgroup level. Although the primary reason for this is that we need to be able to match product associations to customer characteristics, there are other good reasons for using the product subgroup level which we will explain later in this chapter.

Since we use the product subgroup level to describe both customers and products, the initial automatic recommendations are made at the product subgroup level. The final decision about which product item to select is made using a content filtering approach. This offers a number of advantages. As we will

see it enables potentially inappropriate recommendations to be identified (for example, promoting meat to vegetarians) and also enables specific business objectives to be included in the selection process (for example, to promote particular brands or new products). To do this, we use customer segmentation and some heuristic rules (rules of thumb) that can be applied automatically at the end of the recommendation process.

6.2 The data to be used

Having identified the business issue, and outlined what type of solution we require, the *second stage in our generic data mining method* is to decide what data we will need to support this type of system.

In the approach outlined above, each customer is characterized by their NRS at the product subgroup level, rather than at the product group level. This is the same requirement that we had when creating the Simple Customer Level (CLA) model described in 4.2, “The data to be used” on page 49. This requires transaction data and some unique customer identifier to perform the aggregation.

The generation of association rules also requires the same transaction level data, a unique customer identifier, and the product hierarchy information. As we will see, the data mining tool that we use automatically performs any necessary aggregations at the different levels in the product hierarchy. However, we require the raw transaction level data to be available during the mining process.

6.2.1 Data model required

The data model that you require to perform the data mining for a product recommender is a subset of the data model described in 4.2, “The data to be used” on page 49. The minimum data requirement to implement the solution is specified in the following list:

Transaction data

1. Specific customer identifier
2. Date and time of transaction
3. Item purchase (identified by UPI code or equivalent product code)
4. Product price (per item or by unit of measurement)
5. Quantity purchased (number of items or units purchased)

Product data

1. A table matching product code to product name, subgroup code to subgroup name, and product group code to product group name
2. A product taxonomy that links product code to product subgroup code, and product subgroup code to product group code.

Using this information the customer records should be constructed, with one record per customer having the following variables:

1. Specific customer ID
2. NRS in each product subgroup

6.3 Sourcing and preprocessing the data

As in the previous sections, the *third stage in our generic data mining* method requires us to think about the requirements for sourcing and pre-processing the data. Since our requirements here are identical to those for the CLA model we refer you to, “A customer level aggregation (CLA) data model” on page 55 for further details.

6.3.1 Additional considerations

Although the requirements to create a database table for the TLA are identical we also require the raw transaction data as input for the associations mining function. The associations mining only requires the unique customer identifier and the product identifier from each transaction. However, where we need to filter the associations for a defined period of time, it is also necessary to have the data of the transaction.

Because we are now mining all of the transaction level data, this can have a significant impact on your database and before you embark on a full associations mining run, it is advisable to obtain some benchmarks for runtime performance and impact on the database loading. Given a representative sample of the transaction data (say 1%), then the associations mining scales approximately linearly with the number of records you mine. So the benchmarks can be scaled directly based on the number of transactions (~100x benchmark time).

Association rules simply use the fact that at some point during the specified period of time the customer purchased the item identified by the item code, it takes no account of the quantity purchased. While this is not normally an issue you should be aware that the association rules we use make no distinction between different amounts of products purchased.

6.3.2 The example data set

In the description that follows we have again used the example data set described in 4.3, “Sourcing and preprocessing the data” on page 56. Although this is somewhat limited in its scope, it does allow us to understand how the data mining is performed and how the recommendations are generated. In 6.6, “Interpreting the results” on page 162, we will discuss the results obtained by a more comprehensive data set obtained from a trial of the type of personalized product recommender that we describe in the following sections.

6.4 Evaluating the data

Having created and populated our data models, *the fourth stage in our data mining method* is to perform an initial evaluation of the data itself. We usually do this in three steps:

- Visual inspection
- Identifying missing values
- Selecting the best variables

These three steps are described in 4.4, “Evaluating the data” on page 63.

6.5 The mining technique

The *fifth stage in our generic data mining method* is to identify the appropriate data mining techniques that we are going to use and how we are going to apply them. To determine the required relationship between the different products purchased, it is clear that the associations mining function is the appropriate technique to use, but it is not obvious precisely how it should be employed. We also make use of the clustering mining technique that we described in 4.5.1, “Choosing the clustering technique” on page 69.

6.5.1 The associations mining technique

Associations mining is the process that enables you to discover which combinations of products your customers purchase and the relationships that exist at all levels in your product hierarchy. This includes specific rules that tell you, for example, Chateau Hursley1999 (a particularly good red wine) is purchased in combination with Bedfont Stilton (a particularly smelly blue cheese). More general rules from different levels in the product hierarchy are also

generated, for example, German Red Wines are purchased with French Cheese, or at a higher level it simply tells us that Wine and Cheese are purchased together. The relationships discovered by the data mining are expressed as association rules. Association rules take the form:

Left-hand side implies right-hand side.

Traditionally, the technique has been used to perform market basket analysis. In the context of market basket analysis an example association rule may have the following form:

If product A is purchased, then this implies product B will also be purchased at the same time.

In the language of association rules, the left-hand side is called the rule “Body” and the right-hand side, the rule “Head”. In general, the rule Body can contain multiple items, but the rule Head only has one item, for example:

If product A and B and C are purchased, then this implies that product D will also be purchased.

In addition to the rule, the associations mining also calculates some statistics about the rule. Four statistical measures are usually used to define the rule and these are the *Confidence* in the association, the *Support* for the association, the *Lift* value for the association and the *Type* of the association. A definition of these statistics is given in the note below.

Note: In market basket analysis:

Confidence measures the fraction of baskets on the left-hand side that also contain the product on the right-hand side of the rule. If product B is present in 50% of the baskets containing product A, then the Confidence is 0.5. Expressed another way, if we know that product A is in a particular basket, then product B will also be found in the same basket on 50% of occasions.

Support measures the fraction of all baskets that contain both the left-hand side and the right-hand side of the rule. Support therefore measures the fraction of baskets for which the rule is true. If product A and product B are found in 10% of the baskets, then the Support will be 0.1.

Lift is ratio of the rule Confidence to the expected confidence of finding the right-hand side of the rule in any basket. For example, if product B was found in only 5% of all baskets, then the Lift for the rule would have a value of 10.0. What this is saying is that in the 5% of cases where A and B are in the same basket, then the association is occurring ten times more often than would be the case if B were selected by chance. Lift is therefore a measure of how the rule improves our ability to predict the right-hand side of the rule.

Type indicates the statistical significance of the rule as either positive ("+"), negative ("-"), or neutral ("."). A Type of "+" indicates that the rule is statistically significant and has a positive correlation. For example, the rule A=>B with a Type "+" means that the purchase of product A implies that B will be purchased. A Type "-" indicates that the rule is statistically significant, but that there is a negative correlation. For example, the rule A=>B with a Type "-" indicates that the purchase of product A is actually having a negative effect on the purchase of product B. A type of "." indicates that the rule is neutral and that there is no statistical significance in the relationship between product A and product B.

6.5.2 Applying the mining technique

To perform the associations mining we use both the transaction level data and the product hierarchy data to calculate the required association rules.

Generating associations rules

The associations mining technique that we use automatically combines the transaction level data at the customer level by using the customer_identifier as the transaction field. This is equivalent to doing market basket analysis, but instead of using the transaction-identifier to group products purchased during the same transaction, we use the customer-identifier to group all purchases by the same customer over a specified period of time. In other words, we treat all of the products purchased by the customer as if they were in one large basket.

In addition as we have already explained, the product hierarchy data is used by the mining tool to automatically determine associations, not only between individual products, but also between product subgroups, between product subgroups and products, between product groups and subgroups and so on. Associations at all levels in the product taxonomy are therefore derived.

Although the associations mining tool can generate rules with multiple terms in the rule body (for example, [Product A] & [Product B] => [Product E]). In the case of the product recommendation system described in this chapter, only single terms are used. These can be either individual products, product subgroups or product groups.

The format of the rules produced by the mining tool can be presented in a number of ways including the tabular rule display shown in Figure 6-1.

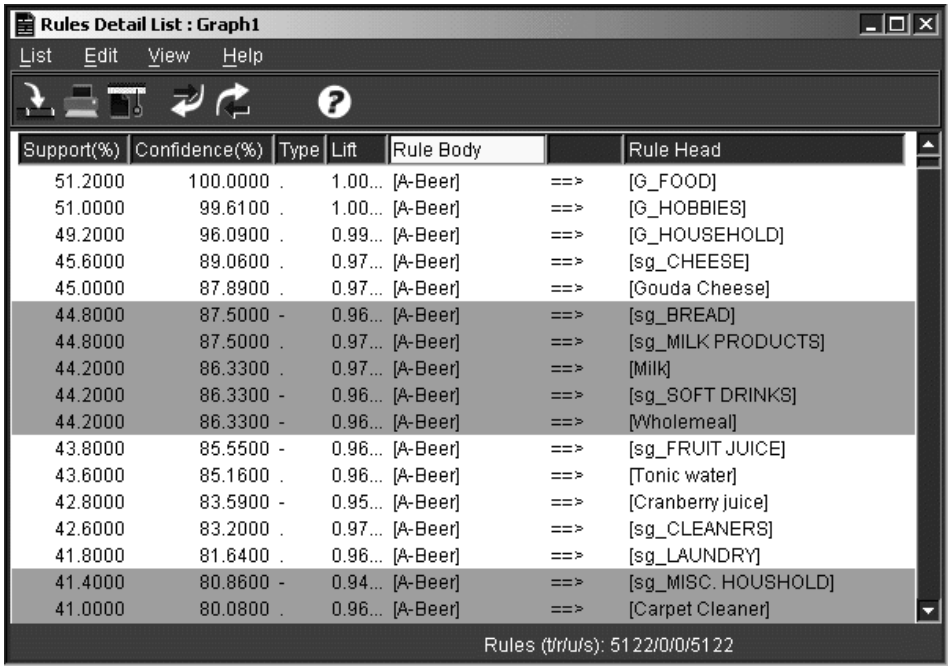


Figure 6-1 Highest confidence association rules for A-Beer

Interpreting association rules

The tabular rule display shows a subset of the rules produced from the example data set. The rules in the display are sorted in alphabetical order on the rule Body, which in this case brings the product A-Beer to the top of the list. The rules have been generated to only include one item in the rule Body which then implies the purchase of the item in the rule Head. Individual products have their product name, for example, Gouda Cheese. If the item is a subgroup, then it is designated with the prefix “sg_”, and the subgroup name is capitalized, for example, sg_CHEESE. If the item is a product group, it is designated with the prefix “G_”, and again the name is capitalized, for example, G_FOOD. The table also includes columns for the “Support”, the “Confidence”, the “Lift” and the “Type” of the rule.

The way to read the list of rules for the purchase of A-Beer is as follows. The first rule in the list simply states that:

> *If a customer purchases the product A-Beer, then they will also purchase products in the product group of food (G_FOOD) with a 100% confidence.*

This is not an altogether surprising conclusion, and the rule is true for 51.2% of customers, indicated by the Support. However, we must note that the Lift value here is 1.0 and the Type is neutral (.), which indicates to us that the purchase of A-Beer has no influence on whether a customer will purchase food or not. Because we are looking for rules that have some predictive value, it is possible to remove all rules from the list that have a neutral Type and this is shown in Figure 6-2.

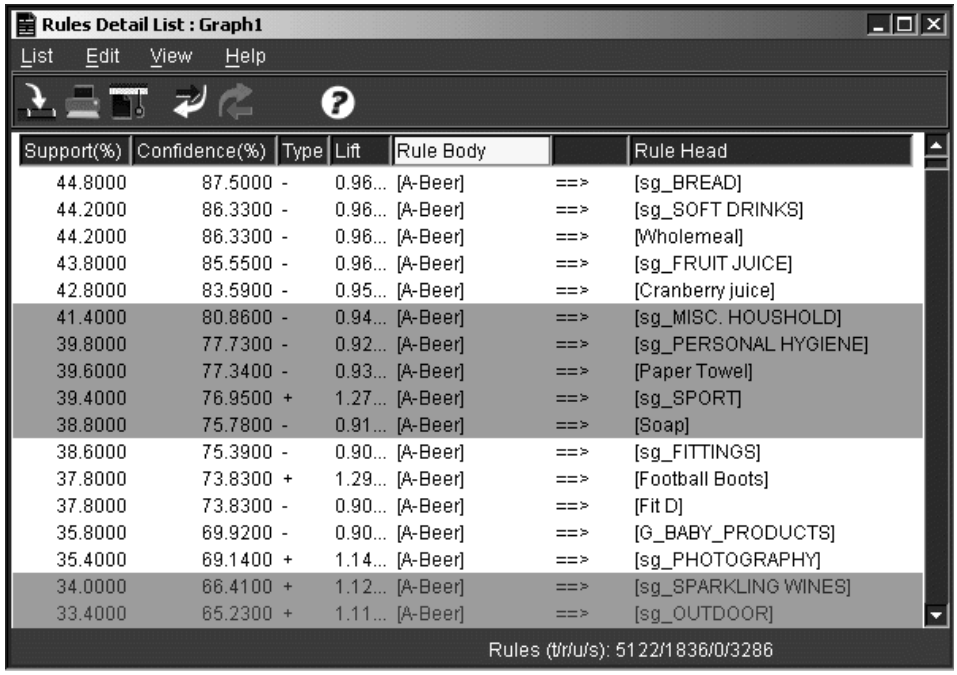


Figure 6-2 Association rules for A-Beer that have some predictive value

This brings to the top of the list a rule which tells us the following:

> *44.8% of customers purchase A-Beer and purchases from the Bread subgroup, but only with an 87.5% confidence.*

This rule has a negative Type (-) and a Lift of less than 1.0 (0.96) which tells us that this is happening less than we would expect, based upon sales of bread in general. Therefore, customers who purchase A-Beer have a tendency not to purchase bread. A similar relationship exists in the next rule between A-Beer and products in the soft drinks subgroup.

The other rules in the list are also very interesting. They tell us, for example, that one of the main reasons why customers who purchase A-Beer purchase less bread than expected, is that they purchase less Wholemeal bread than would be expected. A similar picture is painted between A-Beer and Fruit Juice, in particular Cranberry Juice. This combination seems to paint a picture that drinkers of A-Beer are not particularly health conscious. Coupled with the fact that there is also a reluctance to purchase from the Personal Hygiene products, in particular Soap, this tends to paint a rather unflattering view of the purchasers of A-Beer. This does not mean that negative type rules always imply unflattering attributes, because we may find similar negative rules between people who purchase health products and the product A-Beer. The interpretation that you put on the rules is purely subjective.

The negative Type rules tell us what A-Beer drinkers do not purchase, but what products do A-Beer drinkers favour? To answer this question we can filter the rules again to look for rules with a positive Type (+) as shown in Figure 6-3.

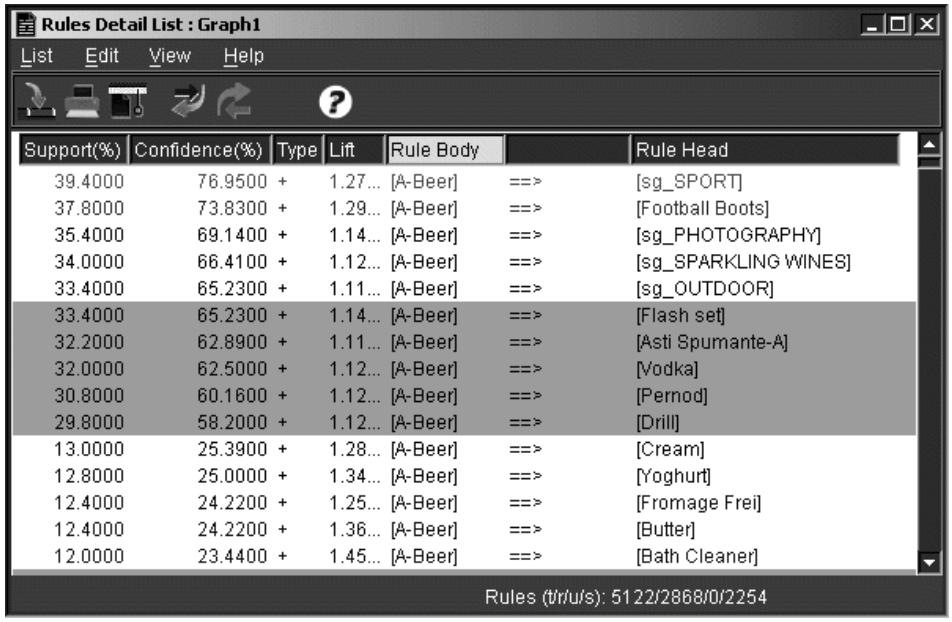


Figure 6-3 Association rules for A-Beer with a “+” Type correlation

We now have a picture that the purchasers of A-Beer now have a preference for Sport, in particular football, for Photography and Outdoor activities and also enjoy other alcoholic products. They also have a liking for some dairy products and purchase Bath Cleaner, so perhaps they are not as bad as we first imagined.

Another way of representing this information is in the form of a graph as shown in Figure 6-4.

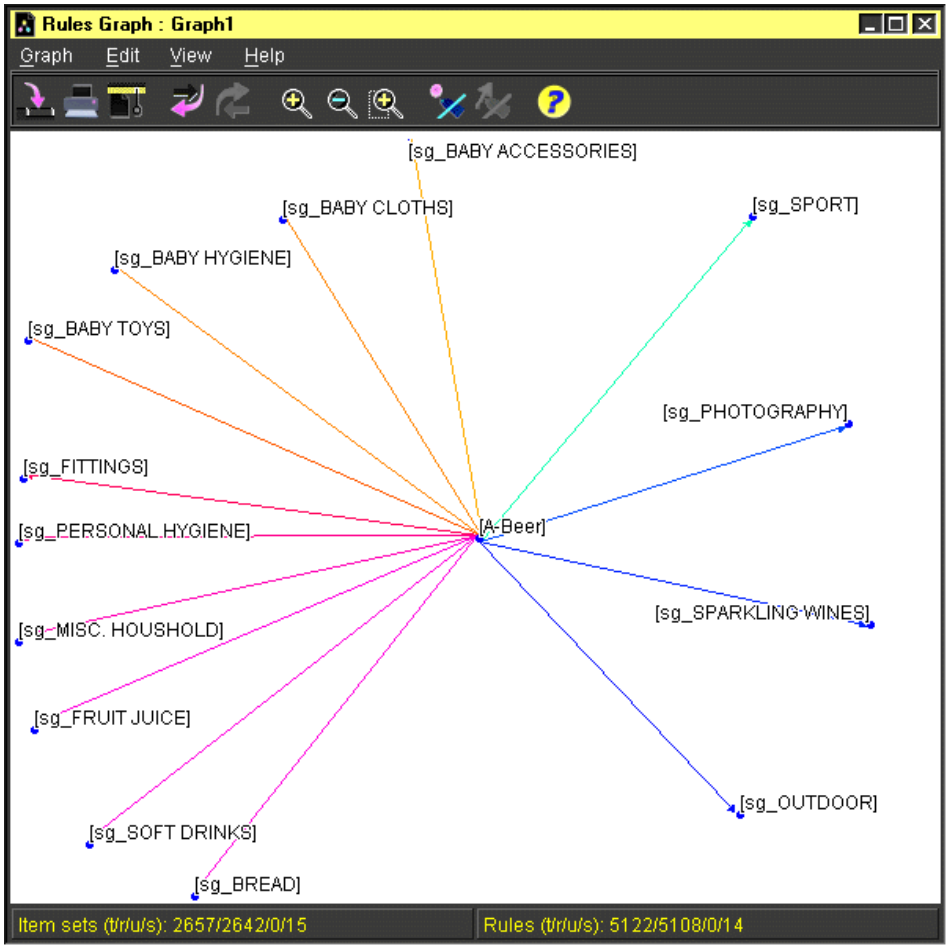


Figure 6-4 Graphical view of association between A-Beer and related subgroups

In this graph the relationship between A-Beer and the related subgroups and groups is shown as a series of arrows linking the different product levels. In the graph we have chosen to represent the shade of the line by rule Lift and have arranged them in a clockwise order with the highest Lift association, sg_SPORT (at the top right), to the lowest Lift association, sg_BABY ACCESSORIES (at the top left). The rules depicted on the right-hand side have “+” type associations and the rules depicted on the left have “-” type associations. This type of graph, when accompanied with the rules and a brief description, is a valuable way of explaining to your marketing department or a brand manager how a particular

product or group of products relate. Filtering on the rule Confidence or Lift or any of the other parameters enables us to construct the relationships that we want to highlight, and there are a variety of ways of showing the same information to bring out different messages.

We can clearly make some very interesting deductions about the relationship of one product to other products. The challenge that faces a large retail organization is that with many hundreds, thousands or even tens of thousands of products, it is not possible to sift through all of the products in this way to determine which other products should be promoted to a particular customer. To do this we need to automate the process.

6.5.3 Using the associations results to compute the product records

In 4.2, “The data to be used” on page 49, we explained that if you want to compare each product to each customer, then you need to represent each product by a record with the same number of fields as there are product subgroups. We identified that what was needed to do this was a measure of the products affinity to the product subgroups and that the associations mining results would enable us to do this.

To assist in understanding what is happening we begin by looking at a customer record. Figure 6-5 shows graphically the customer record for a typical customer selected from “Hobbies” Shopper Type segment that we discussed in Chapter 4. We will call this customer our reference customer. Each customer is characterized by their NRS in each of the product subgroups, and our reference customer’s NRS can be compared to the average NRS for all Hobby Shopper Type customers which is also shown Figure 6-5.

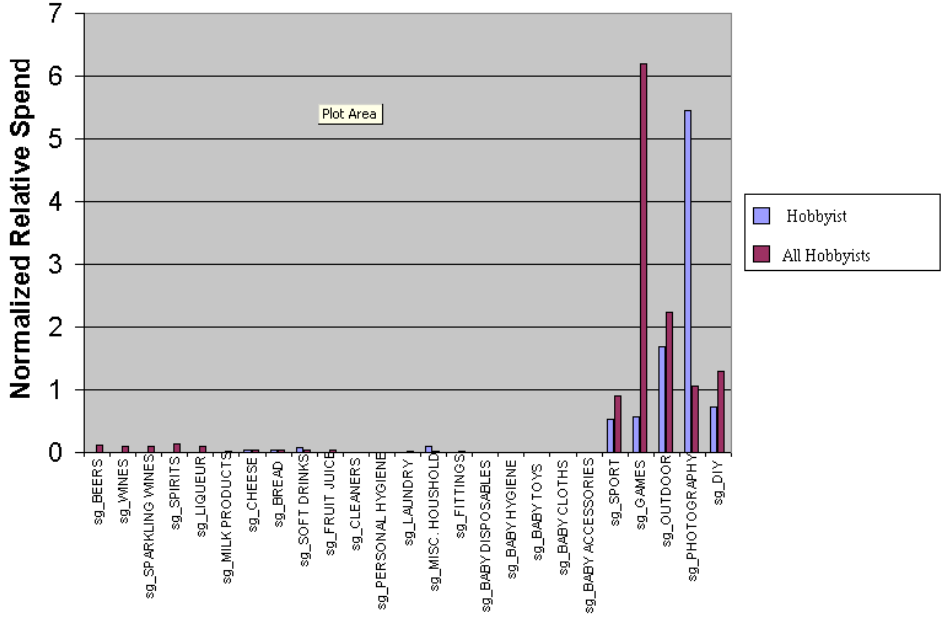


Figure 6-5 Histogram of the normalized relative spend for a customer from the "Hobbyist" segment compared to the average for the whole segment

This type of histogram display immediately shows you that our reference customer has a similar profile to the average for the "Hobbyist" segment, but not exactly. This customer has a preference for photography products, rather than for games products, which is the preferred subgroup for other Hobby Shoppers. This example illustrates that just because a customer is in a particular market segment, it does not mean that there are no important differences in their product preferences. A personalized product recommendation system has to take account of this.

To compare product records with customer records we need first to generate the records and then draw or create a similar display.

Method for generating the product records

This degree of affinity is measured by using an approach very similar to that described in the reference *Personalization of Product Recommendations in Mass Retail Markets*. This approach combines the benefits of the content and collaborative filtering approaches by using heuristic rules to define a value for each component of our product record. These rules take into account both the

relationship between products within the product hierarchy and any association rules that exist at the subgroup and product group levels in the hierarchy. The calculation of the score is relatively simple as you will see by considering the example shown in Figure 6-6.

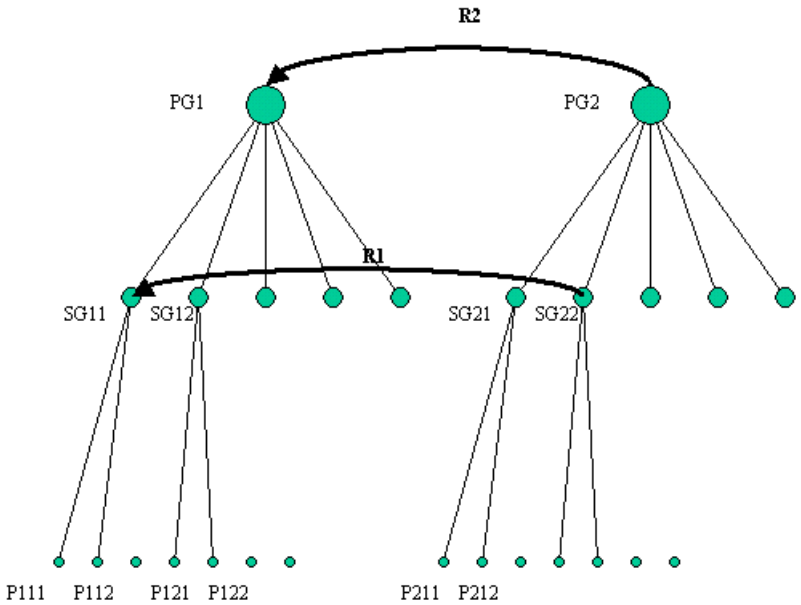


Figure 6-6 Calculating the product vector using product hierarchy and association rules

The figure shows part of a 3 level product hierarchy where we define the product groups by the names PG1, PG2 and so on. The subgroups under each product group have a names SG11, SG12, SG21 and so on, where the first of the two suffixes defines the product group of which it is a subgroup and the second number is the subgroup number within the group. Similarly, products have P111, P112, P121, P122 and so on where the first two numbers defined in the product group and subgroup and the third number is the product identifier itself.

In the simple case illustrated, there are just two association rules (R1 and R2):

Rule 1 shows an association between subgroup SG22 and SG11 (SG22 => SG 11).

Rule 2 shows an association between product groups PG2 and PG1 (PG2 => PG1).

To calculate the product records we have to construct a table with as many columns as there are subgroups and as many rows as there are subgroups. In the case of our example data set this would have 25 rows and columns.

Note: To apply the heuristic rules and compute the scores, perform the following steps:

- Step 1

  For each column read the subgroup name (for example, SG1) and set the cell in the table with the corresponding subgroup name to -1 (all the diagonal values in the table are set to -1). If the row corresponds to a subgroup (for example, SG2) which is in the same product group as the column name (for example, PG1), then give it a score of 0.5 provided the score is not already 1.0.

- Step 2

  For each column get the corresponding product group name (PG1, for example). If there are any association rules from any other product group to this group (PG2 => PG1), set the cell in the table where the row corresponds to a subgroup of the product group on the left-hand side of the rule to 0.25 (for example, for column SG11, then cells with row names SG21, SG22 and so on, would get a score of 0.25). If the score in the cell is already greater than this value, then keep the original value.

- Step 3

  For each column read the subgroup name. If there are any association rules from any other subgroups to this subgroup (SG23 => SG11), set the cell in the table where the row corresponds to a subgroup of the product group on the left-hand side of the rule to 1.0.

It would be possible to devise alternative heuristics that have more dramatic effects, but we have found that using these heuristics tends to generate the product records that maximize the affinity between products that are different, but not too different, from what customers already purchase. We have also made one additional change to the heuristics described in the reference *Personalization of Product Recommendations in Mass Retail Markets*, by setting the diagonal cells to -1, which we do to suppress recommendations in the same subgroup that customers already purchase in, rather than use explicit rules to exclude this possibility.

The result of applying these heuristics of the example data set (shown in Figure 6-6) is shown in Table 6-1.

Table 6-1

	SG11	SG12	SG13	SG21	SG22	SG23
SG11	-1	0.5	0.5	1	0	0
SG12	0.5	-1	0.5	0	0	0

SG13	0.5	0.5	-1	0	0	0
SG21	0.25	0	0	-1	0.5	0.5
SG22	0.25	0	0	0.5	-1	0.5
SG23	0.25	0	0	0.5	0.5	0.5

Each row of this table is now the product record for all of the products within the subgroup corresponding to the row name. Essentially, what we are doing is finding the affinity of the product subgroups with each other.

Method for comparing the customer and product records

To compare a customer record with a product record and to calculate a score for how well they match, you simply multiply the two records together column by column and sum the result. So that we don't bias the result against product records that have a lot of zero values, we also divide each cell by the sum of the values in each record before we multiply.

Note: The customer and product records are compared by treating each record as a vector, calculating the vector dot product and using this as the score.

For every customer, the match with every product subgroup is performed and the scores are then ranked. The product subgroup from which a product can be recommended is then the highest ranking subgroup.

6.5.4 Generating scores using only the product hierarchy

To illustrate what is happening when we multiply customer records and product records and generate recommendation scores, we first look at the case where the product record is generated using only the product hierarchy (only Step 1 in the method for generating product records). In Figure 6-7 you can see the customer record for the reference Hobby Shopper we considered earlier, and an example of the product record for the sports products subgroup (sg_SPORT). Also the product recommendation scores are shown for this customer.

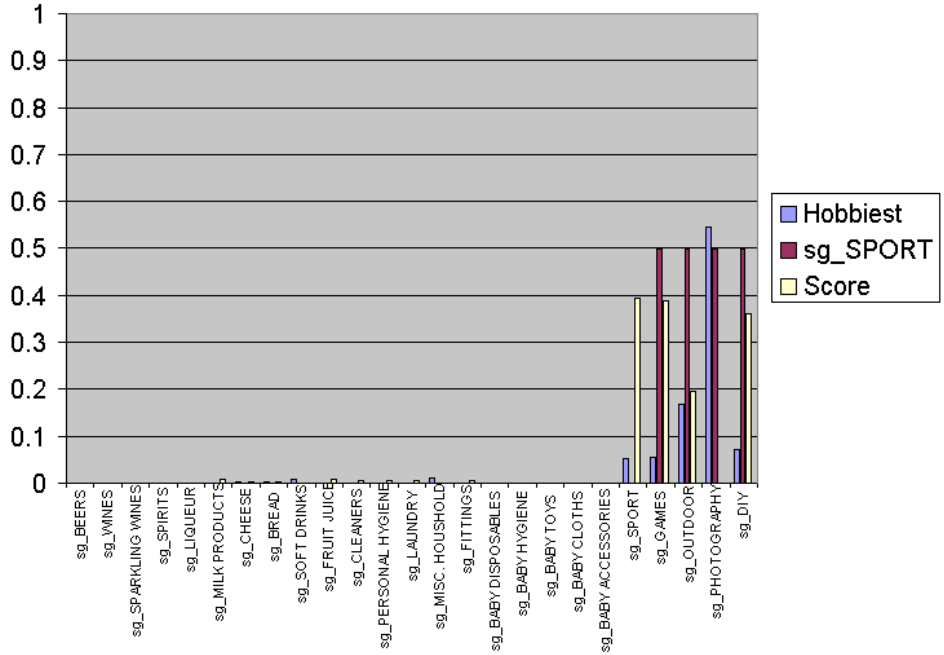


Figure 6-7 Comparison of the customer and product records for the reference hobby shopper together with the recommended scores in each product subgroup using only product hierarchy

Because we have only used the product hierarchy to generate the product records, you can see that the product record for sg_SPORT has a value of 0.5 for all of the subgroups in the Hobbies product group, except sport itself which has a value of -1 (not shown in the figure, because we do not show values below zero). All other subgroups have a value zero. This customer purchases from all of the subgroups in the hobbies category, but primarily from the photography subgroup, and to a lesser extent from the outdoor subgroup. When the two are multiplied you get a correspondingly high score (0.39) for the sg_SPORTS subgroup. There are similar product vectors for each of the hobbies subgroups, but following the multiplication of each, the highest score is still sport followed by sg_GAMES (0.38), sg_DIY (0.35) and sg_OUTDOOR (0.19). It should also be noted that because of the scoring approach we are using to promote cross-sell, the recommendations made favor the subgroup in which the customer purchases the least, which is the desired effect.

The limitations imposed by only using the product hierarchy to generate the product vectors is now very clear. This is just a form of content based filtering and we are only able to offer products that correspond to the customers current buying pattern. At best we are limited to cross-selling products from subgroups in

which the customer normally purchases, or to subgroups within the same product groups from which the customer normally purchases. What we want to be able to do is to expand our opportunities to cross-sell to other product groups and subgroups using the product purchasing behavior of other customers.

6.5.5 Generating scores including association rules

To demonstrate what happens when we include association rules in the generation of the product records, we are going to use the association rules from the example data set. Our method for generating the product record requires association rules between the product groups and between the product subgroups.

To come to the decision about which rules are most appropriate to use is somewhat subjective and we used the following approach. We selected associations rules that contain only product group and subgroup names, and then we disregarded all of the rules with a neutral or negative “Type” and concentrate only on those rules with the highest Lift values. In addition, we rejected all rules with a Support of less than 25% and all rules with a Confidence below 40%.

Our justification for these choices are as follows. Using rules that have the highest Lift is an easy decision, because we require our rules to imply a strong association. One problem that can occur when choosing the higher Lift rules is that you often get high Lift rules, if there are only a few customers who purchase some unusual combinations of products. While these may make interesting suggestions, we decided to play safe and reject rules with a minimum Support value of 25%. This ensures that the rules are representative of a significant proportion of our customer base.

The minimum rule Confidence was also chosen for similar reasons. If the rule Confidence is low, then although high Lift implies that more people than expected purchase this combination, a low rule Confidence tells us that this is not going to be as appealing as a cross-selling opportunity.

The final reason for the choice was that the resulting rules produce an interesting mix of cross-selling suggestions and the values selected are not too dissimilar from those used in 5.6, “Interpreting the results” on page 118, where we have some empirical evidence for successful deployment of this type of recommender.

Using these values, the initial set of associations rules obtained for all customers was reduced from 5122 (with a minimum Support of 25%) to just 260. An example of some of these rules is shown in Figure 6-8.

Figure 6-8 Association rules between product groups and subgroups

Generating the recommendation scores

Using the 260 rules generated from the associations mining, we can generate a complete product records table and use this to score our reference Hobby Shopper. The results are shown in Figure 6-9.

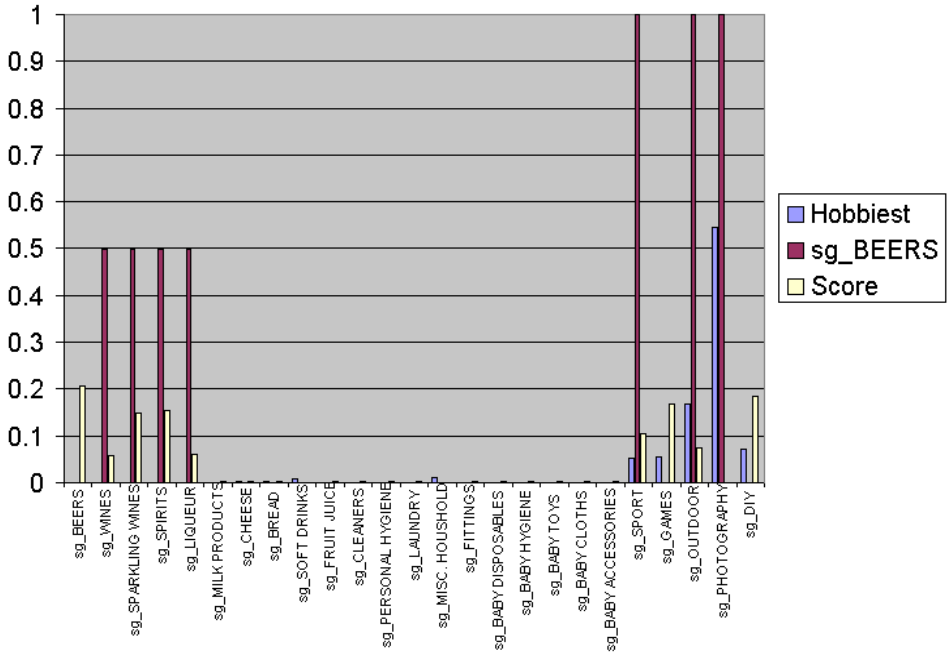


Figure 6-9 Comparison of the customer and product records for the reference Hobby Shopper together with the recommended scores in each product subgroup using product hierarchy and association rules

The effect of introducing the associations rules on the resulting recommendation scores is immediately apparent. The highest score is obtained for the beer subgroup (sg_BEERS; 0.22), followed by the DIY subgroup (sg_DIY; 0.2), and so on. Also the product record for sg_BEERS is shown in Figure 6-9, which illustrates what is happening. Although our reference customer does not purchase products in the Alcohol product group, the associations rules show that there is a strong implication that other customers who purchase in sport, outdoor and photography subgroups will also purchase beer. Because our reference customer has a strong propensity to spend in each of these categories, by multiplying the customer record with the beer product record results in the high score for beer products.

Having seen how the recommendation system works for our target customers, we need to determine what sorts of recommendations are made to the different types of customer that we have. In 6.6, “Interpreting the results” on page 162, we review the overall performance of the recommendation system by looking at the subgroups from which cross-sell recommendations are made to our different shopper type segments. Before doing this however, we need to explain how the actual products that are offered are selected from the recommended subgroups.

6.5.6 Selecting the products to recommend

While it would have been possible to extend the method described above to make recommendations of individual products, rather than at the subgroup level, there are a number of reasons why this is not desirable. The recommendations made are based on identifying associations that are supported by a significant proportion of our customers (we used 25% as a guide). If we attempt to do this at the individual product level, then we will have many more rules, but the rules will have much lower support.

In most retail organizations while the product subgroups (or their equivalent in your organization) remain stable over time, the individual products you sell do not, and therefore developing stable associations rules from historical transaction data becomes a more difficult challenge. It also precludes the promotion of new products and services where there has not been sufficient time to discover meaningful associations. We therefore conclude that the product subgroup level is the most appropriate.

Note: Where your particular organization has a different type of product taxonomy to the one we have described, the appropriate level to use should be selected on the basis of this stability criterion.

We still have to address the issue of precisely which product or products to select from within the subgroup or subgroups that are recommended. This can be done in a number of ways, but the technique that we use now relies more on content filtering rather than collaborative filtering. By this we mean that the products recommended are now selected from the recommended subgroup using the following types of heuristic rules:

- *Select the most popular product purchased from within the subgroup by all customers. This choice can also be appropriately weighted for such factors as profitability or the desire to promote a particular product.*
- *Select the most popular product, or the most unusual product (depending on how radical you want to be) from the recommended product subgroup for people who are "most like" the target customer.*

In the latter case by "most like" we mean those customers who have similar spending profiles. This group is just the group of customers that we identified in Chapter 4, using the data mining segmentation technique. In this case we can select popular or profitable products that customers, who are in the same segment as the target customer, also purchase from the recommended subgroup. A good example of this is described in the reference *Personalization of Product Recommendations in Mass Retail Markets*. In this case when the

recommended product subgroup was CHOCOLATE, the most popular product for the total population of shoppers were Mars Bars. While this was appropriate for many shoppers, the most popular chocolate item for shoppers in the equivalent of our Baby Products shopper type, was Nestle Smarties, a product very popular with small children.

This type of content based selection also has the additional advantage of identifying that a particular recommendation may be inappropriate. As an example, the promotion of meat products to a group of people who are identified as vegetarians may not be a good strategy. In the case where there are no customers, within the same segment, who purchase products from the recommended subgroup, then this may be a cause for concern. In this case we can automatically select products from one of the other recommended subgroups from which customers in this segment do purchase, but flag the missed opportunity to the sales and marketing department for them to determine if this is a new niche opportunity or a special interest group. Using the segmentation mapping technique we described in Chapter 4 would allow this type of analysis to be undertaken.

It is clearly advisable to review the types of recommendations that are going to be made prior to deploying an automatic recommendation system so that inappropriate recommendations can be avoided. Where you have large numbers of customers, this cannot be done by looking at each individual, and some way of summarizing the recommendations that are being made is required. In 6.6, “Interpreting the results” on page 162, we will use this type of technique to interpret the results obtained from our example data set.

6.6 Interpreting the results

We have seen that the personalized product recommender described in the previous section appears to make sensible product recommendations for our reference customer, but what types of recommendations are made to the different groups of customers, and are these viable recommendations, and how do we interpret the results? This is the *sixth stage in our generic data mining method*.

This section looks at the type of recommendations that are made to the 500 customers in our example data set. In the following section, where we look at deploying the mining into our business, we will consider what happens when we apply this technique to larger numbers of customers and to a much wider range of products.

6.6.1 Interpreting the recommendations that were made

To determine if the results of the recommendations that were made are sensible, we need to look at types of cross-sell suggestions that are being made for all of the customers in our example data set. The decision of whether the results are sensible is of course purely subjective and the real test is how our customers will respond to the recommendations we make. However, using business knowledge at this stage will give us confidence that the automatic recommendations are at least appropriate. We can use our judgement to has made, and we can compare the most frequently purchased subgroup for each customer with the recommended subgroup.

To help interpret our results we can use our data mining tool visualizers to show aggregated recommendations. In our first example, shown in Figure 6-10, we show the recommended product subgroup (the one with the highest score) that are made to customers from the Hobby Shopper segment. We also use the product subgroup in which they have their highest expenditure to identify the different types of customer in this segment.

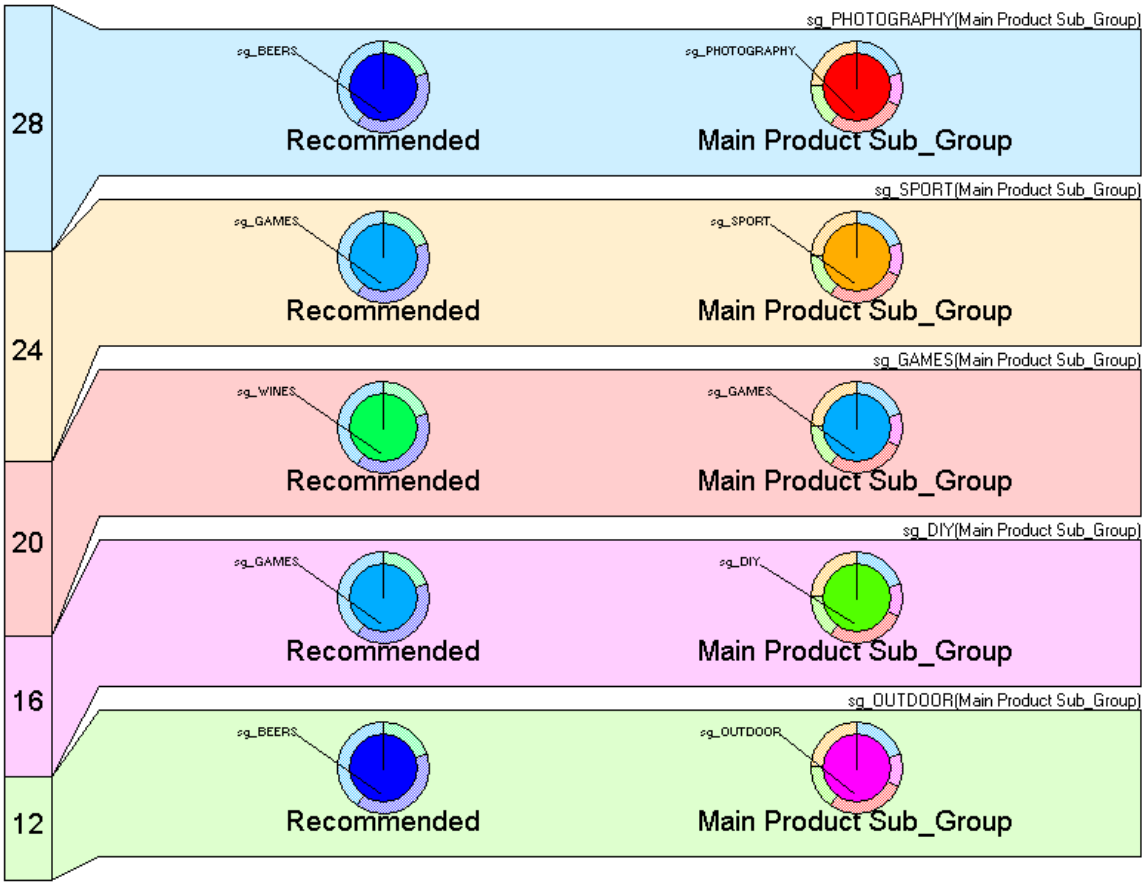


Figure 6-10 Comparison of the product subgroups recommend in comparison the customers preferred product subgroup for the Hobby Shopper Type customers

Because customers in the Hobby Shopper segment have a tendency to purchase predominantly from one category, the use of the Main Products Sub_Group from which they normally purchase gives us a good indication of the type of customer. It is therefore easy to interpret what is driving the cross-sell recommendation. In summary the results show that:

Customers who purchase from the sg_OUTDOOR subgroup of products (hiking boots, maps, compass) or PHOTOGRAPHY subgroup are offered BEER, while people who prefer GAMES are offered WINE. This is a more interesting result when we remember that BEER and WINE are in a subgroup that is only weakly related to customers in this segment as we saw in Figure 6-5 earlier.

Customers who purchase SPORT and DIY products are not offered alcohol, but instead are offered GAMES products from within the same product group. The reason being that there are no strong associations rules between these subgroups and subgroups in other product groups.

However, some customers in the Hobby Shopper segment do purchase alcoholic products, and so when we come to select the appropriate Beer or Wine product to offer, we can use the content filtering approach to make this selection. It is important to realize however that although there are such customers in this segment who purchase alcoholic products, an analysis of Hobby Shopper customers shows that there are no significant association rules to justify the recommendation to purchase Beer. This recommendation has resulted from the behavior of customers in other segments. Therefore, we would not have made the offer without the advantage of collaborative filtering approach.

In the case where our customers have a clear preference for products within a single subgroup (such as, the Hobby Shoppers) this type of analysis gives a clear indication of the relationship between customer preference and the recommendations made. However, where the product preference is less obvious and customers purchase from a range of product subgroups, it is less easy to interpret. In this case it is better to look at the segment as a whole and compare the range of preferred products to the range of recommendations. An example of how this is done is shown in Figure 6-11.

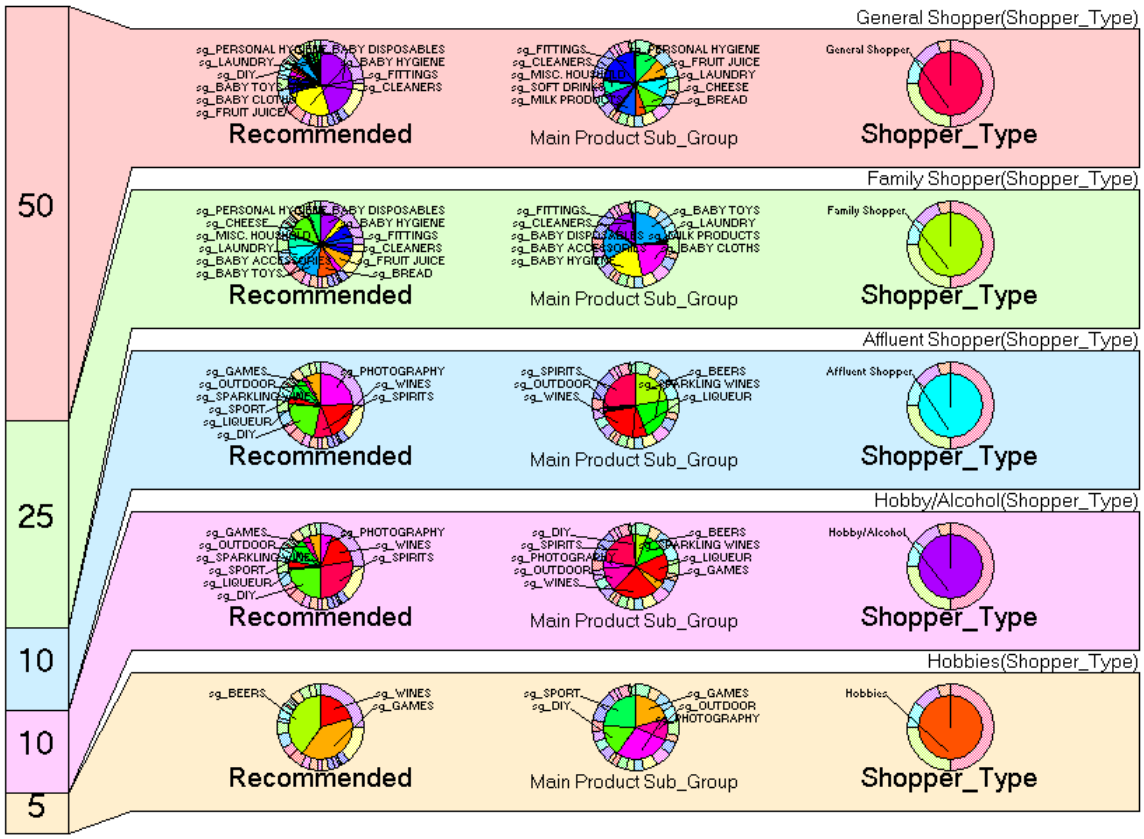


Figure 6-11 Comparison of some of the most frequently purchased subgroups and corresponding to the recommended subgroups for each customer type

The Hobby Shopper Type is shown on the bottom line. We now have an aggregated view of the results from Figure 6-11, which gives us a distribution of the recommendations made. This can be compared with the distribution of the preferred or main product subgroups.

In the case of the other shopper types, for example the Affluent Shopper Type, the recommendations are less clear. Here there is a general tendency to recommend Photography, DIY products, and Wines to a group of customers who show a preference for Outdoor and Sports and Beer. Because without a lot of analysis we don't know the proportion spent in each of the different subgroups, it is difficult to determine precisely what is happening, but that is why we needed an automatic product recommendation system. In general, because the heuristic rules that we used seek to recommend products that are different, but not too different from what customers already purchase, the results are in line with our expectations.

Clearly the example data set that we have used to demonstrate how the technique works is limited and in many ways not indicative of a large scale retail organization. The real question is, does the data mining technique described work, when scaled to a large retail organization with large numbers of customers many tens of thousands of products? Perhaps more importantly, do customers respond positively to the types of recommendations made? We address these questions in the following section.

6.7 Deploying the mining results

The *seventh and final stage in our data mining method* reminds us, we have to be able to deploy our mining solution into our business. In the case of a personalized product recommendation system this is a complex issue to address, because it depends on precisely how your business interacts with its customers. As we mentioned at the beginning of this chapter, there are an increasingly wide range of routes through which you interact with your customers, and recommending products to your customers is just one thing that you have to think about. So in what sort of ways can this be achieved? In this section we are going to answer this question with an example of the sort of thing that can be done.

6.7.1 The typical deployment scenario

The scenario is a supermarket chain (Safeway) in the United Kingdom. In this case Safeway wanted to make a trial of the concept of automatically recommending products to customers who are shopping, using a remote shopping system. In this case supermarket customers were encouraged to use Personal Digital Assistants (PDAs) of the type shown in Figure 6-12 to compose and transmit their orders to the store, which assembles them for subsequent pickup.

Figure 6-12 The type of PDA used in the customer trial

The recommender is meant to provide an alternative source of new ideas for customers who now visit the store less frequently. The recommendations were generated using the data mining technique described in this chapter by matching products to customers, based on the expected appeal of the product and the previous spending of the customer.

A trial of the recommendation system was undertaken using customers from two stores, each with around 20,000 customers. The stores stock in excess of 30,000 different products and it was from this range that recommendations had to be made. The products are organized by the supermarket into a three level hierarchy comprising the 30,000 products organized into 2303 product subgroups and these subgroups into 99 product groups. In the trial, it was determined that not all of the product groups should be eligible for recommendation, and so the system avoided recommending tobacco, health products, and other inappropriate product classes. This is clearly a business decision, but it also highlights the flexibility of the technique to include or exclude products depending on the business objectives.

6.7.2 Evaluating customers' responses to the recommendations

The results obtained during the trial have been reported in the reference *Personalization of Product Recommendations in Mass Retail Markets*, and in this section we summarize the main conclusions from this report.

As we explained at the beginning of the chapter, it is essential that the data used to develop the data mining is representative of the range of customers to whom the products are to be recommended. The data used to develop the recommendations therefore came from a sample of 8000 customers with above-average spending data and comprised eight weeks of product-level transactions data.

The trial was conducted in two phases. During Phase 1 of the trial, using only one of the two stores and an early version of the product recommender, a total of 1,957 complete orders were processed from customers using the PDA system. Of these, 120 orders (6.1%) contained at least one product chosen from the recommendation list. (It is important to note that the recommended list, by design, contained no products previously purchased by this customer.) An objective of the product recommender is to provide a boost in revenue comparable to the spontaneous purchases a shopper may make while walking through the store or after receiving a flyer in the mail. By this measure, the results for the initial recommender were somewhat disappointing: the corresponding boost in revenue was 0.3% over and above the revenue generated by products bought from a personal shopping list that was maintained by the customer.

As the trial program progressed, the development team noticed that the distribution of spending in the different product categories was different for items bought from a recommended list in comparison to personal shopping list that each customer maintained, even though the distribution of items available from each list were quite similar. For example, wines accounted for only 3.5% of the revenue from the main shopping list, but 8.7% of the revenue from the recommended list. By contrast, products in the household care category accounted for 12.1% of the revenue from the main shopping list, but only 4.6% from the recommended list.

The results were interpreted to mean that there are a set of categories in which recommendations are more welcome than others, and a series of interviews with participating customers confirmed that interpretation. Customers wanted more "interesting" recommendations, for example, wines meet that description, but household care products do not. Using this information, which suggested an interesting refinement to the process described in the previous sections, the list of subgroups from which recommendable products were drawn was reduced to emphasize those categories in which the spending percentage from the recommended list exceeded that on the customer shopping list, and

de-emphasizes the others. The aim was to creating a more “fun” or welcome set of recommendable products. A second source of items eligible for recommendation was also included, namely new products introduced within the last month.

In the second phase of the study, the new recommender was introduced into the original store and a second store and a new set of customers were added to the trial. For the original store, the fraction of orders containing at least one recommended product marginally increased from 6.1 to 7.7%, and the revenue boost increased from 0.3 to 0.5%. For the new store however the response was much greater, and 25% of the orders included at least one recommendation, with a revenue boost of 1.8%, a respectable number, given the tight profit margins of the supermarket business. As the investigators pointed out, you get only one chance to make a first impression.

The trial showed that using this type of product recommendation system can encourage shoppers to try new things, but not drastically new things: 51% of the acceptances from the recommended lists corresponded to subgroups in which the shopper had spent no money in the previous three months, but only 4% corresponded to new product groups. This needs to be compared with the statistics for the actual recommendations, where 33% of the recommended products on average were from subclasses the shopper had not spent in before, and 16% were from product classes that were new to the customer. Outside the environment of the recommender, the rate of trying new subgroups was substantially lower, and the rate of trying new product classes is practically zero.

The results obtained clearly show that the type of personal recommendation system described in this chapter can make an impression on your retail business.

7

The value of DB2 Intelligent Miner For Data

Throughout this book we have been concentrating on how data mining can be used to address specific business issues, rather than concentrating on the detailed capability of the data mining tool that we use. It should come as no surprise to you that we have been using the IBM DB2 Intelligent Miner For Data product (IM for Data) to perform all of the data mining functions described in the previous chapters. Although we have used some of the data mining techniques that are available to us, IM for Data offers a number of other tools and techniques that we have not been able to describe how to use.

In this chapter we take the opportunity to explain the advantages and benefits of using IM for Data by presenting an overview of where the product is placed in relation to its competitors and also a summary of all of its functions and capabilities.

The recent introduction of DB2 Intelligent Miner Scoring (IM Scoring) opens up a whole range of new possibilities for deploying the results of data mining into your business process. You may be concerned that IM Scoring is only applicable to IM for Data and DB2 databases. This is not the case, and we provide some further details on the background to IM Scoring.

7.1 What benefits does IM for Data offer?

The IBM mining products offer a full range of data mining techniques and capabilities that can be used to address a wide range of business problems. The techniques included support the identification of unknown patterns and trends within your stored data. If you use DB2 as your database, then IM for Data is the only mining product that is fully integrated with the DB2 database. If you do not use DB2, then you can still use IM for Data through the Relational Connect feature in DB2 for read only access or through a companion product called DataJoiner. DataJoiner provides additional capabilities that give federated access to a number of different data sources in read and write mode. So even if your data is distributed across a number of databases and file systems you can still perform data mining.

Some straightforward benefits are:

- High performance and scalability of mining functions. Therefore, there is no need for sampling.
- Customized front-ends and visualizers.
- Excellent clustering algorithm for demographic data. Best suited for detection of niche clusters.
- Highly efficient association algorithm and visualizer. Only association algorithm which supports taxonomies.
- Mines any known database format via Relational Connect feature (part of IBM DB2 Universal Database) or IBM DataJoiner.
- Coupling with SAP Business Information Warehouse.
- Real time model deployment through IM Scoring.
- Running on parallel machines.

IM for Data is one of the industry leaders in integration of data mining and database technology.

7.2 Overview of IM for Data

IM for Data provides the user with the complete spectrum of state-of-the-art data mining algorithms, together with a range of tools for data preparation and statistical analysis. The data mining algorithms are categorized as follows:

- Clustering
- Associations discovery
- Sequential patterns discovery
- Classification

- Prediction
- Similar time sequence discovery

Further details of each of these algorithms are given in the following sections.

Each data mining algorithm can be considered as a tool that is used to perform a specific discovery task (that means using the mining analogy — discovering seams of information). Continuing the mining analogy, the tools can also be used in concert to perform more specific complex operations and to extract diamonds or nuggets of valuable information. Next, we describe the different components and functions of the product and how this is achieved.

7.2.1 Data preparation functions

The data to be mined can be in the form of flat files or held in tables within a relational database. If the database option is used, then the preferred database is the IBM DB2 Universal Database (DB2 UDB). Connection to other databases is possible using the DB2 Relational Connect feature to accede in read-only mode to ORACLE, SYBASE, SQLServer or using IBM DataJoiner product to accede in read and write mode to INFORMIX, ORACLE, SYBASE, TERADATA, SQLServer.

Once the desired input data has been selected, it is usually necessary to perform certain transformations on the data. IM for Data provides a wide range of data preparation functions that help to quickly transform the data to be analyzed. The data preparation functions of IM for Data are:

- Aggregate values: To aggregate values of existing fields, for example, monthly salary to annual salary.
- Calculate values: To create new fields with the result of a calculation of existing fields. The Calculate values preprocessing function creates new fields using SQL expressions. These new fields are appended to the input data to create the output data.
- Clean up data sources: To delete database tables or views in the database, usually those no longer used as input or output data.
- Convert to lower-case or upper-case: To convert one or more fields in the output data.
- Copy records to file: To copy records from a database table or view to a flat file (you can also sort by the field you specify).
- Discard records with missing values: To remove input data records containing a missing (NULL) value in any of the fields you specify.
- Discretize into quantiles: To assign input data records to the number of quantiles you specify.

- Discretize using ranges: To assign input data records by splitting the value range of a continuous field into intervals, and then mapping each interval to a discrete value.
- Encode missing values: To encode missing values in the input data by specifying one or more fields to search for missing values and further specifying which value to use as a replacement for any missing values in these fields.
- Encode non-valid values: To encode values found in the first input data if they do not match valid values from the second input data and further specifying which value to use as a replacement for any such value.
- Filter fields: To filter the input data fields to get an output table or view containing only the fields you specify or the ones you didn't specify.
- Filter records: To filter the input data records to get only the records you specify for which a given condition is true.
- Filter records using a value set: To compare field values in a first input data with values in a value set specified for a second input data; then filter the records whose input field contains a value present in the value set.
- Get random sample: To reduce input data to a smaller sample by specifying the size of the sample as a percentage of the input data.
- Group records: To summarize groups of records into a single record that contains aggregated values of the group.
- Join data sources: To join two database tables or views based on one or more pairs of join fields from the input data.
- Map values: To map values found in the first input data to values found in the second input data.
- Pivot fields to records: To split each record of the input data into multiple records.
- Run SQL statements: To submit SQL statements.

Data preparation functions are performed through the GUI, reducing the time and complexity of data mining operations. The user can transform variables, input missing values, and create new fields through the touch of a button. This automation of the most typical data preparation tasks is aimed at improving productivity by eliminating the need for programming specialized routines.

7.2.2 Statistical functions

Although the data mining tools are designed to discover information from the data, understanding the data structure in terms of outlying values or highly correlated features is often necessary if the full power of the mining techniques is to be realized. Therefore after transforming the data, the next stage is usually to analyze it. IM for Data provides a range of statistical functions to facilitate the analysis and the preparation of data, as well as providing forecasting capabilities. For example, you can apply statistical functions like regression to understand hidden relationships in the data, or use factor analysis to reduce the number of input variables. The statistical functions included are:

- Factor analysis: Discovers the relationships among many variables in terms of a few underlying, but unobservable, quantities called factors.
- Linear regression: Used to determine the best linear relationship between the dependent variable and one or more independent variables.
- Polynomial regression: Used to determine the best polynomial relationship between the dependent variable and one or more independent variables.
- Principal component analysis: Used to rotate a coordinate system so that the axes better match the data distribution. The data can be now described with fewer dimensions (axes) than before.
- Univariate curve fitting: Finds a mathematical function that closely describes the distribution of your data.
- Univariate and bivariate statistics: Descriptive statistics, especially means, variances, medians, quantiles, and so on.

7.2.3 Mining functions

All of the mining functions can be customized using two levels of expertise. Users who are not experts can accept the defaults and suppress advanced settings. However, experienced users who want to fine tune their application are provided with the capability to customize all settings according to their requirements. It is also possible to define the mode in which the data mining model will be performed. The possible modes are:

- Training mode: In which a mining function builds a model based on the selected input data.
- Test mode: In which a mining function uses new data with known results to verify that the model created in training mode produces adequate results.
- Application mode: In which a mining function uses a model created in training mode to predict the specified field for every record in the new input data.

The user can also use data mining functions to analyze or prepare the data for a further mining run. The following sections describe each mining algorithm in more detail, using typical commercial examples to illustrate the functionality.

Clustering

Clustering is used to segment a database into subsets, the clusters, with the members of each cluster having similar properties. IM for Data can perform clustering by using either a statistical clustering algorithm (Demographic Clustering) or a neural network algorithm (Kohonen Clustering), depending on the type of the input data set. The neural clustering algorithm requires the user to specify the number of clusters required; the statistical clustering algorithm automatically determines the “natural” number of clusters.

When clustering is performed there are no preconceived notions of what patterns exist within the data; it is a discovery process. The results of the clustering process can be visualized (see 4.6, “Interpreting the results” on page 79) to determine the composition of each cluster. Visualization graphically presents the statistical distributions of the characteristics of those records that compose the cluster in comparison with the data set as a whole. Tabular output is also provided to enable further analysis.

In addition to producing graphical and tabular output, a “cluster model” is also generated (Training Mode). It is also possible to generate a user-defined table, which can include selected information from the input records, together with the cluster number of the segment to which the record has been assigned. The output table can also include details on the next nearest cluster and a measure of the confidence in the degree of matching to the nearest and next nearest clusters for each record (Test Mode). An Application Mode is also provided, in which new data records are assigned to clusters and an output table generated.

In the commercial environment clustering is used in the areas of cross-marketing, cross-selling, customizing marketing plans for different customer types, deciding on media approach, understanding shopping goals, and so forth.

Associations

The association algorithm, developed at the IBM Almaden Research Center in San Jose, California, compares lists of records to determine if common patterns occur across the different lists. In a typical commercial application the algorithm looks for patterns such as whether, when a customer buys paint, they also buy paintbrushes. More specifically, it assigns probabilities; for example, if a

customer buys paint, there is a 20% chance that they will buy a paintbrush. The advantage of this approach is that it compares all possible associations. It also finds multiple associations, for example, if a customer buys paint and paint brushes, there is a 40% chance they will also buy paint thinner.

When the algorithm runs, it potentially creates hundreds or thousands of such rules. The user can however select a subset of rules that have either higher confidence levels (a high likelihood of B given A) or support levels (the percent of transactions in the database that follow the rule) or high lift (the ratio of measured to expected confidence for a rule). It is up to the user to read the rules and decide if the rules are:

- Chance correlations (for example, paint and hair rollers were on sale the same day and therefore were correlated by chance).
- Known correlations (for example, the paint and paint brush correlation is something that would have been known).
- Unknown but trivial correlations (for example, red gloss paint and red non gloss paint correlation may be something unknown, and is unimportant to know).
- Unknown and important correlations (for example, paint and basketballs, which may be something previously unknown and very useful in both organization of advertising and product placement within the store).

Association discovery is used in market basket analysis, item placement planning, promotional sales planning, and so forth.

The association algorithm also includes the capability to include a taxonomy for the items in the lists (for example, paint and a paintbrush are hardware) and the algorithm will discover associations across the taxonomy (for example, there is a 50% confidence that customers who buy hardware also buy soft furnishing).

Sequential patterns

The purpose of discovering sequential patterns is to find predictable patterns of behavior over a period of time. This means that a certain behavior at a given time is likely to produce another behavior or a sequence of behaviors within a certain time frame.

The rule generation method is a variation of the association technique. It analyzes the shopping behavior of customers, for example, over time. Instead of looking at 10,000 purchases, the algorithm looks at 10,000 sets of purchases. These sets are, for example, lists of purchases from a sequence of shopping trips by a single customer. As a typical commercial example, one set of lists may be the purchases of computer:

- Computer in December

- Computer games and joy stick in January
- Additional computer memory and larger hard drive in March

If this sequence, possibly with different time scales but the same order, were repeated across a number of customers, then the sequential association algorithm would typically return a rule, such as:

> If following the purchase of a computer, the customer purchases computer games, then there is a 30% chance that extra computer memory will be purchased in a subsequent visit to the store.

The algorithm also includes the capability to define minimum and maximum time periods between the items in the lists. This would, for example, enable the above rule to include the statement that computer memory will be purchased no earlier than one month and within three months of the purchase of the computer games.

Sequential pattern detection can therefore be used to discover associations over time. This is especially useful in commercial applications, such as direct marketing, or the design special advertising supplements, and so on.

Classification

Classification is the process of automatically creating a model of classes from a set of records that contain class labels. The induced model consists of patterns, essentially generalizations over the records that are useful for distinguishing the classes. Once a model is induced, it can be used to automatically predict the class of other unclassified records. IM for Data has two classification algorithms, a tree induction algorithm (modified CART regression tree) and a neural network algorithm (back propagation), to compute the classes.

The tree and neural network algorithms develop arbitrary accuracy. While neural networks often produce the most accurate classifications, trees are easy to understand and modify and the model developed can be expressed as a set of decision rules.

Commercial applications of classification include credit card scoring, ranking of customers for directed mailing, and attrition prediction. One of the main uses of the tree algorithm is to determine the rules that describe the differences between the clusters generated by the clustering algorithm. This is achieved by taking the output table from the clustering algorithm and constructing the decision tree using the cluster label as the class.

Value prediction

Value prediction is similar to classification; the goal is to build a data model as a generalization of the records. However, the difference is that the target is not a class membership but a continuous value, or ranking. IM for Data has two prediction algorithms: a neural network algorithm and a Radial Basis Functions (RBF) algorithm. The radial basis function is particularly efficient and is appropriate for value prediction with very large data sets.

Similar time sequences

The purpose of this process is to discover all occurrences of similar subsequences in a database of time sequences. Given a database of time sequences, the goal is to find sequences similar to a given one, or find all occurrences of similar sequences. The powerful alternatives afforded by multiple methods are enhanced by the fact that several of the methods are supported by more than one mining technique. Multiple techniques are often used in combination to address a specific business problem.

7.2.4 Creating and visualizing the results

Information that has been created using statistical or mining functions can be saved for further analysis in the form of result objects. The result objects can be visualized using a variety of graphical displays or the results exported to spreadsheets (for example, EXCEL, LOTUS 123), or to browsers (for example, Netscape, Explorer), or to specific statistical packages (for example, SPSS).

Result objects can be used in several ways:

- To visualize or access the results of a mining or statistical function
- To determine what resulting information you want to write to an output data object
- To be used as input data, when running a mining function in test mode to validate the predictive model representation by the result
- To be used as input data, when running a mining function in application mode to apply the model to new data

7.3 DB2 Intelligent Miner Scoring

DB2 Intelligent Miner Scoring (IM Scoring) is an economical and easy-to-use mining deployment capability. It enables users to incorporate analytic mining into Business Intelligence, eCommerce and OLTP applications. Applications score records (segment, classify or rank the subject of those records) based on a set of predetermined criteria expressed in a data mining model.

These applications can better serve business and consumer users alike — to provide more informed recommendations, to alter a process based on past behavior, to build more efficiencies into the online experience; to, in general, be more responsive to the specific situation at hand. All scoring functions offered by the DB2 Intelligent Miner for Data are supported.

The IM Scoring is an add-on service to DB2, consisting of a set of User Defined Types (UDTs) and User Defined Functions (UDFs), which extends the capabilities of DB2 to include some data mining functions. Mining models continue to be built using the IM for Data, but the mining application mode functions are integrated into DB2. Using the IM Scoring UDFs, you can import certain types of mining models into a DB2 table and apply the models to data within DB2. The results of applying the model are referred to as scoring results and differ in content according to the type of model applied. The IM Scoring includes UDFs to retrieve the values of scoring results.

The results of applying the model are referred to as scoring results and differ in content according to the type of model applied. The IM Scoring includes functions to retrieve the values of scoring results.

The IM Scoring is available on the following operating systems:

- AIX
- Solaris
- Windows NT, Windows 2000
- Linux, Linux/390

Summary of functionality

The application mode for the following IM for Data mining and statistical functions are supported by the IM Scoring:

- Demographic and neural clustering
- Tree and neural classification
- RBF and neural prediction
- Polynomial regression

Scoring functions are provided to work with each of these types. Each scoring function includes different algorithms to deal with the different mining functions included within a type, for example, the clustering type includes demographic and neural clustering and so, scoring functions for clustering include algorithms for demographic and neural clustering. For all the supported mining functions, you build and store the model using the IM for Data. Models must then be exported to an external file.

Exchanging models

In support of facilitating the exchange of mining models between applications, IM Scoring makes full use of the Predictive Model Markup Language (PMML) published by Data Mining Group.

PMML is a standard format. Based on the Extensible Markup Language (XML), it provides a standard by which data mining models can be shared between the applications of different vendors. It provides a vendor-independent method of defining models so that proprietary issues and incompatibilities are no longer a barrier to the exchange of models between applications. You can find more information about PMML on the Web site of the Data Mining Group, at:

`http://www.dmg.org`

The IM Scoring includes a facility for converting models, built using IM for Data, to the PMML format. Using this facility, you can select the PMML format when you export the model from the IM for Data GUI. Conversion to PMML is not necessary when importing models into DB2 using the IM Scoring functions. The model import functions read models in either PMML or IM for Data format.

Using the PMML standard allows the models created by IM for Data to be used in databases other than DB2, and IM Scoring also supports ORACLE Cartridge Extenders.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 184.

- *Intelligent Miner For Data Applications Guide,* SG24-5252
- *Intelligent Miner For Data: Enhance Your Business Intelligence,* SG24-5422
- *Getting Started with Data Warehouse and Business Intelligence,* SG24-5415
- *Mining Relational and NonRelational Data with IBM Intelligent Miner For Data Using Oracle, SPSS, and SAS As Sample Data Sources,* SG24-5278

Other resources

These IBM publications are also relevant:

- *Using the Intelligent Miner for Data V6.1,* SH12-6394
- *Intelligent Miner for Data V6.1 Using the Associations Visualizer,* SH12-6396
- *Intelligent Miner Scoring, Administration, and Programming for DB2,* SH12-6719

These external publications are also relevant as further information sources:

- *Data Preparation For Data Mining.* Dorian Pyle. Morgan Kaufmann Publishers, March 1999. ISBN: 1558605290
- *Data Mining Your Website*. Jesus Mena. Digital Press, July 1999. ISBN: 1555582222
- *Intelligent Data Analysis: An Introduction.* Michael Berthold and David J. Hand. Springer Verlag, September 1999. ISBN: 3540658084
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation.* Ian H.Witten and Eibe Frank. Morgan Kaufmann Publishers, October 1999. ISBN: 1558605525

- *Mastering Data Mining: The Art and Science of Customer Relationship Management.* Michael J. A. Berry and Gordon Linoff. John Wiley & Sons, December 1999. ISBN: 0471331236
- *Data Mining Techniques: For Marketing, Sales, and Customer Support.* Michael J. A. Berry and Gordon Linoff. John Wiley & Sons, May 1997. ISBN: 0471179809
- *Advances in Knowledge Discovery and Data Mining.* Usama M. Fayyad, et al. MIT Press, March 1996. ISBN: 0262560976
- *Multivariate Statistical Analysis; A Conceptual Introduction.* Sam Kash Kachigan. 2nd ed. Radius Press, June 1991. ISBN: 0942154916
- *Personalization of Product Recommendations in Mass Retail Markets.* R.D. Lawrence, et al. Yorktown Heights, New York: IBM T. J. Watson Research Center, November 1999.

Referenced Web sites

These Web sites are also relevant as further information sources:

- `http://www.kdnuggets.com/`
  The Knowledge Discovery Mine Web site that contains a guide to commercial and public domain data mining tools, a newsletter, links to Web sites, and research materials
- `http://www.ibm.com/software/`
  IBM software home page
- `http://www.ibm.com/software/data/iminer/fordata/`
  IBM DB2 Intelligent Miner For Data Web site
- `http://www.ibm.com/software/data/`
  IBM Database and Data Management home page
- `http://www.dmg.org`
  The Data Mining Group Web site

How to get IBM Redbooks

Search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become Redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others.

Glossary

A

adaptive connection. A numeric weight used to describe the strength of the connection between two processing units in a neural network. Values typically range from zero to one, or -0.5 to +0.5.

aggregate. To summarize data in a field.

application programming interface (API). A functional interface supplied by the operating system or a separate licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.

architecture. The number of processing units in the input, output, and hidden layer of a neural network. The number of units in the input and output layers is calculated from the mining data and input parameters. An intelligent data mining agent calculates the number of hidden layers and the number of processing units in those hidden layers.

associations. The relationship of items in a transaction in such a way that items imply the presence of other items in the same transaction.

attributes. or variable or field.Characteristics or properties that can be controlled, usually to obtain a required appearance. For example, color is an attribute of a line. In object-oriented programming, a data element defined within a class.

B

back-propagation. A general-purpose neural network named for the method used to adjust weights while learning data patterns. The classification -neural function uses such a network.

bucket. One of the bars in a bar chart showing the frequency of a specific value.

C

categorical values. Nonnumeric data represented by character strings: for example colors.

chi-square test. A test to check whether two variables are statistically dependent or not. Chi-square is calculated by subtracting the expected frequencies (imaginary values) from the observed frequencies (actual values). The expected frequencies represent the values that were be expected if the variable question was statistically independent.

classification. Assignment of objects into groups based on their characteristics.

cluster. A group of records with similar characteristics.

clustering. A mining function that creates groups of data records within the input data on the basis of similar characteristics. Each group is called a cluster.Assignment of objects into groups based on their characteristics.

confidence factor. Indicates the strength or the reliability of the associations detected.

continuous field. A field that can have any floating point number as its value.

D

database view. An alternative representation of data from one or more database tables. A view can include all or some of the columns contained in the database table or tables on which it is defined.

data format. There are different kinds of data formats, for example, database tables, database views, pipes, or flat files.

data type. There are different kinds of IM For Data data types, for example, discrete numeric, discrete nonnumeric, binary, or continuous.

discrete. Pertaining to data that consists of distinct elements such as character or to physical quantities having a finite number of distinctly recognizable values.

discretization. The act of transforming a set of continuous values in a set of discrete values.

F

field. Or variable or attribute. A set of one or more related data items grouped for processing. In this document, with regard to database tables and views, field is synonymous with column in a database table.

K

Kohonen Feature Map. A neural network model comprised of processing units arranged in an input layer and output layer. All processors in the input layer are connected to each processor in the output layer by an adaptive connection. The learning algorithm used involves competition between units for each input pattern and the declaration of a winning unit. Used in neural clustering to partition data into similar record groups.

L

large item sets. The total volume of items above the specified support factor returned by the Associations mining function.

learning algorithm. The set of well-defined rules used during the training process to adjust the connection weights of a neural network. The criteria and methods used to adjust the weights define the different learning algorithms.

M

metadata. Data that describes data objects.

mining. Synonym for analyzing or searching patterns in data.

model. A specific type of neural network and its associated learning algorithm. Examples include the Kohonen Feature Map and back propagation.

N

neural network. A collection of processing units and adaptive connections that is designed to perform a specific processing function.

NRS. Normalized Relative Spend.

P

prediction. The dependency and the variation of one field's value within a record on the other fields within the same record. A profile is then generated that can predict a value for the particular field in a new record of the same form, based on its other field values.

R

radial basis function (RBF). In data mining functions, radial basis functions are used to predict values. They represent functions of the distance or the radius from a particular point. They are used to build up approximations to more complicated functions.

record. A set of one or more related data items grouped for processing. In reference to a database table, record is synonymous with row.

region. (Sub)set of records with similar characteristics in their active fields. Regions are used to visualize a prediction result.

rule. A clause in the form head <== body. It specifies that the head is true if the body is true.

rule body. Represents the specified input data for a mining function.

rule group. Covers all rules containing the same items in different variations.

rule head. Represents the derived items detected by the Associations mining function.

S

scaling. To adjust the representation of a quantity by a factor in order to bring its range within prescribed limits.

self-organizing feature map. See *Kohonen Feature Map*.

sensitivity analysis report. An output from the Classification - Neural mining function that shows which input fields are relevant to the classification decision.

sequential patterns. Inter transaction patterns such that the presence of one set of items is followed by another set of items in a database of transactions over a period of time.

similar (time) sequences. Occurrences of similar sequences in a database of time sequences.

Structured Query Language (SQL). An established set of statements used to manage information stored in a database. By using these statements, users can add, delete, or update information in a table, request information through a query, and display results in a report.

support factor. Indicates the occurrence of the detected association rules and sequential patterns based on the input data.

T

taxonomy. Represents a hierarchy or a lattice of associations between the item categories of an item. These associations are called taxonomy relations.

taxonomy relation. The hierarchical associations between the item categories you defined for an item. A taxonomy relation consists of a child item category and a parent item category.

trained network. A neural network containing connection weights that have been adjusted by a learning algorithm. A trained network can be considered a virtual processor: it transforms inputs to outputs.

transaction. A set of items or events that are linked by a common key value, for example, the articles (items) bought by a customer (customer number) on a particular date (transaction identifier). In this example, the customer number represents the key value.

transaction id. The identifier for a transaction, for example, the date of a transaction.

V

vector. A quantity usually characterized by an ordered set of numbers.

W

weight. The numeric value of an adaptive connection representing the strength of the connection between two processing units in a neural network.

Index

S

T

U

V

W

IBM Redbooks

Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data

(0.2”spine)
0.17”<->0.473”
90<->249 pages

Mining Your Own Business in Retail

Using DB2 Intelligent Miner for Data

Exploring the retail business issues

Addressing the issues through mining algorithms

Interpreting and deploying the results

The new challenge of integrated solutions is to get more knowledge from data in order to build the most valuable solutions. This IBM Redbook is a solution guide to address the business issues in retail by real usage experience and to position the value of DB2 Intelligent Miner For Data in a Business Intelligence architecture.

Typical retail issues are addressed in this redbook, such as:

How can I characterize my customers from the mix of products that they purchase? How can I decide which products to recommend to my customers? How can I categorize my customers and identify new potential customers?

This book also describes a data mining method to ensure that the optimum results are obtained. It details for each business issue:

- What common data model to use
- How to source the data
- How to evaluate the model
- What data mining technique to use
- How to interpret the results
- How to deploy the model

Business users who want to know the payback on their organization when using the DB2 Intelligent Miner For Data solution should read the sections about the business issues, how to interpret the results, and how to deploy the model in the enterprise.

Implementers who want to start using mining techniques should read the sections about how to define the common data model to use, how to source the data, and how to choose the data mining techniques.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6271-00 ISBN 0738422940